Ray Kurzweil

# Cómo crear una mente

## El secreto del pensamiento humano

lola books

RAY KURZWEIL

# Cómo crear una mente

## El secreto del pensamiento humano

Con prólogo
de José Luis Cordeiro

Traducido del inglés
por Carlos García Hernández

lola books

Copyright © Lola Books GbR, Berlin 2013
www.lolabooks.eu
Copyright © de la traducción: Carlos García Hernández

Queda totalmente prohibida cualquier forma de reproducción, distribución, comunicación pública o transformación total o parcial de esta obra sin el permiso escrito de los titulares de explotación.

Título original:
*How to create a mind: the secret of human thought revealed*
Copyright © Ray Kurzweil, 2012
Publicado con el consentimiento de Ray Kurzweil
Todos los derechos reservados

Diseño de cubierta: Christine Wenning
Copyright © Christine Wenning
Impreso en Clausen & Bosse, Leck
ISBN 978-3-944203-05-8
eISBN 978-3-944203-17-1

Primera edición 2013

# Contenido

Prólogo

Agradecimientos

Introducción

CAPÍTULO UNO
Experimentos mentales históricos

CAPÍTULO DOS
Experimentos mentales sobre el pensamiento

CAPÍTULO TRES
Un modelo del neocórtex: la teoría de la mente basada en el reconocimiento de patrones

CAPÍTULO CUATRO
El neocórtex biológico

CAPÍTULO CINCO
El cerebro antiguo

CAPÍTULO SEIS
Capacidades transcendentes

CAPÍTULO SIETE
El neocórtex digital de inspiración biológica

CAPÍTULO OCHO

La mente como ordenador

CAPÍTULO NUEVE
Experimentos mentales sobre la mente

CAPÍTULO DIEZ
La ley de los rendimientos acelerados aplicada al cerebro

CAPÍTULO ONCE
Objeciones

Epílogo

Notas

Índice

# Prólogo

Ray Kurzweil es probablemente el futurista más influyente del mundo en estos momentos. Pero mucho más allá de futurista, Ray es ingeniero, inventor, empresario, músico, educador y escritor. Todos sus libros sobre tecnología han sido *best-sellers*, comenzando con *The Age of Intelligent Machines* publicado en 1990, *The Age of Spiritual Machines* en 1999, y *The Singularity is Near* en 2005. Ray lanzó su nuevo libro *How to Create a Mind* en diciembre de 2012, y la presente edición en español *Cómo crear una mente* en octubre de 2013. Cada libro ha sido impactante en su momento y todos han tenido importantes predicciones sobre grandes eventos y posibilidades a futuro.

Ray tiene una historia personal impresionante, llena de logros, invenciones y visiones sobre el futuro. Sus invenciones van desde diferentes tipos de escáneres y sintetizadores musicales hasta máquinas de lectura para ciegos y aparatos para reconocimiento de voz. Ray cuenta que desde que tenía 5 años quería ser inventor, y en 1998 fue condecorado como el Inventor del Año por el Instituto Tecnológico de Massachusetts (MIT), su alma mater. Ray ha recibido honores de tres presidentes de los Estados Unidos de América, una veintena de doctorados *Honoris causa* de diversas universidades alrededor del mundo, la Medalla Nacional de Tecnología e Innovación de Estados Unidos y su nombre está en el Salón de la Fama de Inventores de Estados Unidos (National Inventors Hall of Fame). Para ayudar a preparar a las nuevas generaciones, Ray fundó Singularity University junto con Peter Diamandis en el Parque Tecnológico de NASA Ames en el Valle del Silicio (el famoso *Silicon Valley* de California). Por si fuera poco, después de publicar *How to Create a Mind*, Ray ha sido nombrado Vicepresidente de Ingeniería en Google, donde se enfoca en el procesamiento de lenguajes naturales, combinando lingüística computacional e inteligencia artificial.

He seguido las ideas de Ray durante tres décadas y tengo el placer de conocerle personalmente desde hace casi dos décadas. En los últimos 5 años he tenido el privilegio de trabajar con él como uno de los asesores y profesores fundadores de Singularity University, y también colaboré en la revisión de sus predicciones para el año 2009. De hecho, la trayectoria histórica de sus predicciones es impresionante, con una precisión cercana al 90%. Entre sus predicciones más famosas están la caída de la Unión Soviética, la aparición de Internet y el desarrollo de un ordenador o computador capaz de ganar al campeón mundial de ajedrez. En este último caso, Ray

fue incluso conservador pues predijo que esto ocurriría en 1998; cuando Deep Blue de IBM ganó a Garry Kasparov en 1997.

*How to Create a Mind*, el más reciente *best-seller* de Ray en la lista de *The New York Times* , explica los impresionantes avances de la inteligencia artificial y cómo dentro de unos pocos años podremos terminar la ingeniería inversa del cerebro humano. Ray sostiene que para el año 2029 una inteligencia artificial pasará la llamada Prueba o Test de Turing, basado en la idea del científico inglés Alan Turing para saber si un humano es capaz de diferenciar si está escribiendo o hablando con otro humano o con una máquina. Ray incluso explica que la inteligencia artificial tendrá en realidad que bajar su nivel para no ser fácilmente identificada como superior a la inteligencia humana. Eventualmente, si una máquina se comporta en todos los aspectos como inteligente, incluso más inteligente que los humanos, entonces debe ser inteligente.

Ray comienza su nuevo libro con una serie de experimentos mentales para comprender mejor cómo pensamos los humanos. Luego presenta un modelo del neocórtex y plantea su *Teoría de la Mente basada en el Reconocimiento de Patrones* (PRTM, del inglés *Pattern Recognition Theory of Mind*). Ray continúa con un análisis de las diferentes partes biológicas del cerebro y su evolución, para entonces discutir cómo sería un neocórtex digital, creado gracias al crecimiento acelerado de la tecnología.

*Cómo crear una mente* defiende que la mente es una "propiedad emergente" del cerebro, de manera que la creación de cerebros digitales resultará en la creación de mentes digitales. De hecho, el cerebro, actual sustrato biológico de la mente humana, puede ser sustancialmente mejorado gracias a sustratos no biológicos cuidadosamente diseñados y mucho más avanzados. Como diría el gran futurista inglés Sir Arthur C. Clarke, los humanos somos simplemente bípedos con un sustrato basado en carbono (en inglés: *carbon-based bipeds)*. Lo importante no es el sustrato, biológico o no, sino la mente, y las mentes aumentadas gracias a las nuevas tecnologías superarán a las actuales mentes humanas no mejoradas. Ray no sólo considera que la mente es una consecuencia directa del cerebro, sino que además las inteligencias artificiales tendrán conciencia, libre albedrío y hasta identidad propia.Este nuevo libro de Ray es quizá su mejor y más importante obra, como así fue referido por el pionero de la inteligencia artificial Marvin Minsky del MIT, ya que trata específicamente sobre el cerebro humano, que es la estructura más compleja del universo conocido. Tal vez mañana aparezca algún extraterrestre con un cerebro más avanzado que el nuestro, pero hasta entonces el cerebro humano es la estructura más compleja que conocemos. De cualquier forma, el cerebro humano tampoco es tan complejo y en los próximos años podremos imitarlo, simularlo y superarlo con ingeniería inversa gracias a los avances científicos y las tecnologías exponenciales.

A pesar de la gran complejidad del cerebro humano, con sus cien mil millones de neuronas conectadas a través de billones y billones de sinapsis, Ray indica que su "objetivo con este libro no es en absoluto añadir una nueva cita a las millones que ya existen y que atestiguan lo complejo que es el cerebro, sino más bien impresionarle a usted con el poder de su simplicidad. Esto lo realizaré describiendo cómo un ingenioso mecanismo básico que se repite cientos de millones de veces y que sirve para reconocer, recordar y predecir un patrón es el responsable de la gran diversidad de nuestro pensamiento."

Para quienes todavía no creen que una inteligencia menor pueda evolucionar hacia una inteligencia mayor, nosotros mismos somos la prueba de que sí es posible. Hace millones de años los humanos evolucionamos de ancestros simios menos inteligentes, los cuales a su vez evolucionaron de otros mamíferos primitivos todavía menos inteligentes (aunque inteligencia quizá no sea la palabra correcta en este sentido). Hasta ahora, nuestros cerebros biológicos han sido el resultado de la evolución biológica al azar, con resultados buenos y malos, aleatoriamente. En el futuro, los cerebros digitales que vamos a producir serán diseñados, y dichos cerebros artificiales no serán el resultado fortuito de la evolución biológica sino creaciones inteligentes gracias a nuestra evolución tecnológica.

En los pocos meses que han trascurrido desde que la edición inglesa de *How to Create a Mind* apareció en diciembre de 2012, dos proyectos trascendentales sobre el cerebro han comenzado. Por un lado está el Proyecto Cerebro Humano, un esfuerzo médico-científico y tecnológico financiado por la Unión Europea y dirigido por el científico surafricano Henry Makram desde la Escuela Politécnica Federal de Lausana, Suiza. Con un presupuesto de por lo menos mil millones de euros durante los próximos diez años, el Proyecto Cerebro Humano busca simular el cerebro con supercomputadores para reproducir tecnológicamente las características del cerebro humano. Por otro lado está la Iniciativa BRAIN (del inglés *Brain Research through Advancing Innovative Neurotechnologies*), anunciada por el presidente estadounidense Barack Obama con el objetivo de hacer un mapa de cada neurona del cerebro humano. La Iniciativa BRAIN está basada en el exitoso Proyecto Genoma Humano y se prevén inversiones de más de 300 millones de dólares al año durante toda una década.

Las investigaciones multimillonarias que están comenzando tanto con el Proyecto Cerebro Humano en Europa como con la Iniciativa BRAIN en Estados Unidos tendrán resultados impresionantes durante la próxima década. Por si fuera poco, Japón sigue con sus investigaciones avanzadas en el Instituto RIKEN del Cerebro, y China, Rusia y otros países también tienen programas importantes sobre el cerebro, neurociencia e inteligencia artificial. Grandes compañías tecnológicas como Amazon, Apple, Ericsson, Facebook, Google, IBM, Microsoft, Nokia, Samsung y

Sony, por ejemplo, también tienen ya algunos productos y más proyectos en áreas similares. Además existen nuevos *start-ups* que igualmente están trabajando en aspectos fundamentales del cerebro y la inteligencia, tanto natural como artificial. Con todo este interés a nivel nacional e internacional, público y privado, para comprender y mejorar el cerebro humano, yo no tengo la menor duda de que vamos a descubrir cosas maravillosas en los próximos años.

Ray explica que su meta es "comprender de forma precisa cómo funciona el cerebro y luego utilizar el desvelamiento de dichos métodos para comprendernos mejor a nosotros mismos, así como para reparar el cerebro cuando sea necesario y (lo que es lo más importante para este libro) para crear máquinas cada vez más inteligentes." Según Ray, continuaremos fusionándonos con nuestra tecnología en una civilización humano-máquina cada vez más avanzada.

Aunque algunas ideas puedan parecer ciencia ficción, es bueno recordar que muchas veces la ciencia ficción de hoy se convierte en la ciencia real de mañana. La ciencia continuamente abre nuevas puertas y oportunidades al conocimiento humano. De hecho, lo que antes parecía imposible, puede volverse realidad más tarde. Los primeros teléfonos fijos, los automóviles, los aviones, los antibióticos, los satélites artificiales, los ordenadores o computadores, Internet, los teléfonos móviles o celulares, todos parecían magia en su momento. Ahora, afortunadamente, cada uno de esos descubrimientos e invenciones son considerados normales por las nuevas generaciones. Efectivamente, a veces las ideas avanzan de la ciencia ficción hacia la ciencia real. Sir Arthur C. Clarke, un ingeniero que es más conocido como autor de ciencia ficción, escribió hace medio siglo sus famosas tres leyes del futuro:

1. Cuando un científico viejo y distinguido afirma que algo es posible, es casi seguro que está en lo correcto. Cuando afirma que algo es imposible, es muy probable que esté equivocado.

2. La única manera de descubrir los límites de lo posible es aventurarse más allá de ellos, hacia lo imposible.

3. Cualquier tecnología suficientemente avanzada no se diferencia de la magia.

En pocas palabras, lo que hoy puede parecer magia, pronto quizá podría ser realidad. Desde el punto de vista computacional, ya estamos comenzando a reproducir la complejidad del cerebro humano. De hecho, como estima Ray, es posible que una inteligencia artificial pase el Test de Turing en el año 2029 (aunque probablemente sea antes, como demostraron los rápidos avances de dos ordenadores o computadores de IBM: Deep Blue en 1997 y Watson en 2011). Entonces será imposible diferenciar entre una inteligencia artificial y una inteligencia humana.

Poco después las inteligencias artificiales seguirán mejorando y superarán a las inteligencias humanas no modificadas. En el camino, la mayor parte de la humanidad seguirá utilizando la tecnología para aumentar sus capacidades, como hemos hecho hasta ahora (desde lentes hasta prótesis). Luego será posible subir todos nuestros conocimientos, recuerdos, experiencias, amores y hasta sentimientos a ordenadores o computadores (a Internet o a la "nube") que incluso tendrán una memoria expandible y muy superior a la memoria humana actual. La memoria artificial además continuará mejorando y creciendo, al igual que la capacidad y la velocidad de procesamiento de la inteligencia artificial. Todo será parte de un proceso acelerado de mejora de la inteligencia humana gracias a la continua evolución tecnológica.

La humanidad apenas está comenzando el fascinante camino de la evolución biológica a la evolución tecnológica, una nueva evolución consciente e inteligente. Según explica Ray, un kilogramo de "computronio" tiene la capacidad teórica para procesar cerca de $5 \times 10^{50}$ operaciones por segundo, comparado con un cerebro humano que puede procesar entre $10^{16}$ y $10^{19}$ operaciones por segundo (según diferentes estimaciones). De forma que todavía tenemos un potencial enorme por delante, de muchos órdenes de magnitud, para seguir aumentado la inteligencia humana y luego posthumana, pasando de nuestros cerebros bilógicos no mejorados a cerebros post-biológicos aumentados. Como concluye Ray: "nuestro destino es despertar al universo para luego decidir inteligentemente cuál es su futuro imbuyéndole de inteligencia humana en su forma no biológica."

José Luis Cordeiro, MBA, PhD
(www.cordeiro.org)

Director, Nodo Venezuela, The Millennium Project
(www.Millennium-Project.org)

Profesor Fundador, Singularity University, NASA Ames, Silicon Valley, California
(www.SingularityU.org)

Co-fundador, Asociación Transhumanista Iberoamericana
(www.TransHumanismo.org)

Fundador, Sociedad Mundial del Futuro Venezuela
(www.FuturoVenezuela.net)

*A Leo Oscar Kurzweil.*
*Te estás adentrando en un mundo extraordinario.*

# Agradecimientos

Me gustaría expresar mi gratitud a mi esposa Sonia por su amorosa paciencia durante las vicisitudes del proceso creativo, a mis hijos Ethan y Amy, a mi nuera Rebeca, a mi hermana Enid y a mi nieto Leo por su amor e inspiración.

También a mi madre Hannah por apoyar mis primeras ideas e inventos, lo cual me dio la libertad de poder experimentar a una edad temprana, y por mantener vivo a mi padre durante su larga enfermedad.

También a mi editor en Viking, Rick Kot, por su liderazgo y por su constante y perspicaz orientación, así como por ser un experto de la edición.

Gracias a Loretta Barrett, mi agente literario desde hace veinte años, por guiarme astuta y entusiásticamente.

A Aaron Kleiner, mi socio empresarial de toda la vida, por su denodada colaboración durante los últimos cuarenta años.

A Amara Angelica por su denodada y excepcional ayuda investigadora.

A Sarah Black por sus extraordinarias indicaciones en cuanto a la investigación y por sus ideas.

A Laksman Frank por sus excelentes ilustraciones.

A Sarah Reed por su entusiasta apoyo organizativo.

A Nanda Barker-Hook por su experta organización de mis compromisos públicos relacionados con este y otros temas.

A Amy Kurzweil por su orientación en el oficio de escribir.

A Cindy Mason por su ayuda investigadora y por sus ideas sobre la IA y la conexión cuerpo-mente.

A Dileep George por sus perspicaces ideas y por nuestras reveladoras discusiones por email y por otros medios.

A Martine Rothblatt por su dedicación a todas las tecnologías que expongo en este libro y por nuestras colaboraciones en el desarrollo de tecnologías en estas áreas.

Al equipo de KurzweilAI.net, que me ha proporcionado un importante apoyo investigador y logístico para llevar adelante este proyecto. El equipo incluye a: Aaron Kleiner, Amara Angelica, Bob Beal, Casey Beal, Celia Black-Brooks, Cindy Mason, Denise Scutellaro, Joan Walsh, Giulio Prisco, Ken Linde, Laksman Frank, Maria Ellis, Nanda Barker-Hook, Sandi Dube, Sarah Black, Sarah Brangan y Sarah Reed.

Gracias al entregado equipo de Viking Penguin por todo su atento buen hacer. El equipo de Viking Penguin incluye a: Clare Ferraro (presidente), Carolyn Coleburn (director publicitario), Yen Cheong y Langan Kingsley (publicistas), Nancy Sheppard (directora de márquetin), Bruce Giffords (editor de la producción), Kyle Davis (asistente editorial), Fabiana Van Arsdell (directora de producción), Roland Ottewell (editor de copias), Daniel Lagin (diseñador) y Julia Thomas (diseñadora de la cubierta).

Gracias a mis colegas de la Singularity University por sus ideas, entusiasmo y energía emprendedora.

A mis colegas que me han proporcionado inspiración para las ideas que aparecen en este volumen. Quiero incluir aquí a: Barry Ptolemy, Ben Goertzel, David Dalrymple, Dileep George, Felicia Ptolemy, Francis Ganong, George Gilder, Larry Janowitch, Laura Deming, Lloyd Watts, Martine Rothblatt, Marvin Minsky, Mickey Singer, Peter Diamandis, Raj Reddy, Terry Grossman, Tomaso Poggio y Vlad Sejnoha.

Gracias a mis lectores especializados, que incluyen a Ben Goertzel, Davis Gamez, Dean Kamen, Dileep George, Douglas Katz, Harry George, Lloyd Watts, Martine Rothblatt, Marvin Minsky, Paul Linsay, Rafael Reif, Raj Reddy, Randal Koene, Dr. Stephen Wolfram y Tomaso Poggio.

A mis lectores en general y de mi entorno cuyos nombres aparecen más arriba y finalmente gracias a todos los pensadores creativos del mundo que me inspiran cada día.

# Cómo crear una mente

# Introducción

*El cerebro— es más amplio que el cielo—*
*Ya que— puestos el uno al lado del otro—*
*El primero contiene al segundo*
*Con facilidad— y contigo incluido—*
*El cerebro es más profundo que el mar—*
*Ya que— comparando los azules—*
*El uno absorbe al otro—*
*Como las esponjas hacen con los cubos—*
*El Cerebro pesa lo mismo que Dios—*
*Ya que— comparados libra a libra—*
*Se diferenciarán— si acaso—*
*Como una sílaba de su sonido*

—EMILY DICKINSON

Al tratarse del fenómeno más importante del universo, la inteligencia es capaz de transcender las limitaciones naturales y de transformar el mundo según su propia imagen. Bajo nuestro mando, la inteligencia nos ha permitido superar las restricciones de nuestra herencia genética y que al mismo tiempo nos transformáramos nosotros mismos durante el proceso. Somos la única especie capaz de hacer esto.

El relato de la inteligencia humana comienza con un universo que es capaz de codificar información. Este ha sido el factor que ha hecho posible que la evolución tenga lugar. Cómo el universo acabó siendo de la manera que es, es un relato interesante. El modelo estándar de la física tiene docenas de constantes que tienen que ser precisamente las que son. De otra manera los átomos no habrían sido posibles y no habría habido estrellas, ni planetas, ni cerebros, ni libros sobre el cerebro. El hecho de que las leyes de la física estén tan precisamente ajustadas como para permitir la evolución de la información es algo que parece increíblemente poco probable. Sin embargo, el principio antrópico nos dice que si no fuera así, no estaríamos preguntándonos sobre ello. Donde algunos ven una mano divina, otros ven un multiverso. En este multiverso es donde diferentes universos evolucionan y en donde los universos aburridos (los que no contienen información) se extinguen. Pero, independientemente de cómo nuestro universo acabó siendo como es, podemos empezar nuestro relato con un mundo basado en la información.

La historia del universo se desarrolla mediante niveles de abstracción creciente. Los átomos (especialmente los de carbono, ya que pueden crear estructuras ricas en

información mediante enlaces en cuatro direcciones diferentes) formaron moléculas cada vez más complejas. El resultado fue que la física dio lugar a la química.

Mil millones de años después, surgió una molécula compleja llamada ADN que podía codificar con precisión largas cadenas de información y generar los organismos descritos mediante dichos "programas". El resultado fue que la química dio lugar a la biología.

A un ritmo cada vez mayor, los organismos desarrollaron redes de comunicación y de decisión llamadas sistemas nerviosos que podían coordinar las cada vez más complejas partes de sus cuerpos, así como los comportamientos que facilitaban su supervivencia. Las neuronas, que componían los sistemas nerviosos, se juntaron en cerebros capaces de comportarse cada más vez inteligentemente. Así fue como la biología dio lugar a la neurología, ya que los cerebros se encontraban en la vanguardia del almacenamiento y manipulación de la información. De este modo pasamos de átomos a moléculas, de moléculas a ADN, y de ADN a cerebros. El siguiente paso fue exclusivamente humano.

El cerebro de los mamíferos posee una clara capacidad no encontrada en ninguna otra clase de animal: somos capaces de pensar *jerárquicamente*, de comprender una estructura compuesta de diferentes elementos ordenados según un patrón, de representar dicho ordenamiento mediante un símbolo y luego utilizar dicho símbolo como elemento de una configuración todavía más elaborada. Esta capacidad reside en una estructura cerebral llamada neocórtex. En los humanos, el neocórtex ha alcanzado un nivel de sofisticación y capacidad tal que estos patrones han merecido el nombre de *ideas*. Así, mediante un interminable proceso recursivo, somos capaces de construir ideas cada vez más complejas. A esta enorme matriz de ideas recursivamente unidas la llamamos conocimiento. Solo el homo sapiens posee una base de conocimientos que evoluciona, crece exponencialmente y es transmitida de generación en generación.

Nuestros cerebros dieron lugar a otro nivel de abstracción, ya que, además de la inteligencia de nuestros cerebros, contamos con otro factor determinante: un apéndice oponible (el pulgar) que nos permite manipular el medio y construir herramientas. Dichas herramientas representaron una nueva forma de evolución, y así la neurología dio lugar a la tecnología. Solo gracias a nuestras herramientas nuestra base de conocimientos ha podido crecer sin límite.

Nuestra primera invención fue el relato: el lenguaje hablado que nos permitió representar ideas mediante vocablos diferenciados. Con la posterior invención del lenguaje escrito, desarrollamos diferentes formas de simbolizar nuestras ideas. Así, mediante ideas recursivamente estructuradas, las librerías de lenguaje escrito aumentaron enormemente la capacidad de nuestros desprovistos cerebros para retener y expandir nuestra base de conocimientos.

Existe cierta controversia sobre si otras especies, tales como los chimpancés,

poseen la capacidad de expresar ideas jerárquicas mediante el lenguaje. Los chimpancés son capaces de aprender un conjunto limitado de símbolos pertenecientes al lenguaje de signos, lo cual les permite comunicarse con sus instructores humanos. Sin embargo, es evidente que existen claros límites en cuanto a la complejidad de las estructuras del conocimiento con las que los chimpancés son capaces de manejarse. Las frases que pueden expresar se limitan a secuencias simples y específicas de sujeto-predicado, y no son capaces de lograr la indefinida expansión de la complejidad que caracteriza a los humanos. A modo de divertido ejemplo sobre la complejidad del lenguaje generado por los humanos, basta con leer una de las espectaculares frases de varias páginas escritas por Gabriel García Márquez en sus relatos o novelas. Su relato de seis páginas "El último viaje del buque fantasma" se compone de una sola frase y funciona bastante bien tanto en español como en la traducción inglesa (1).

La idea principal de mis tres libros anteriores sobre la tecnología (*The Age of Intelligent Machines*, escrito en la década de 1980 y publicado en 1989, *The Age of Spiritual Machines*, escrito entre mediados y finales de la década de 1990, y *La Singularidad está cerca**, escrito a principios de la década de 2000 y publicado en 2005) es que cualquier proceso evolutivo sufre una aceleración intrínseca (debido a sus cada vez mayores niveles de abstracción) y que además la complejidad y capacidad de los productos nacidos de estos procesos crece exponencialmente. A este fenómeno le llamo ley de los rendimientos acelerados (LOAR)** y concierne tanto a la evolución biológica como a la tecnológica. El ejemplo más claro de la LOAR viene representado por el hecho de que el crecimiento exponencial de la capacidad y de la relación rendimiento/precio de las tecnologías de la información es extraordinariamente predecible. Por eso el proceso evolutivo de la tecnología desembocó inevitablemente en el ordenador, que a su vez ha supuesto una enorme expansión de nuestra base de conocimientos y ha permitido aumentar extraordinariamente la capacidad de comunicación entre las diferentes áreas de conocimiento. La propia web es un buen y poderoso ejemplo de la capacidad de un sistema jerárquico para abarcar una gran variedad de conocimientos y a la vez mantener su estructura intrínseca. En definitiva, podríamos afirmar que el mundo en sí es intrínsecamente jerárquico (los árboles contienen ramas, las ramas contienen hojas, las hojas contienen vetas. Los edificios contienen pisos, los pisos contienen habitaciones, las habitaciones contienen entradas, ventanas, paredes y suelos).

También hemos desarrollado herramientas que en estos momentos nos están permitiendo comprender nuestra propia biología en precisos términos informativos. Además, estamos aplicando sin demora la ingeniería inversa a los procesos de la información que subyacen tras la biología, incluyendo la biología de nuestros cerebros. En forma de genoma humano, ya estamos en posesión del código objeto de

la vida, un logro que en sí fue un extraordinario ejemplo de crecimiento exponencial, ya que la cantidad de información genética que se secuencia en el mundo ha venido más o menos doblándose anualmente durante los últimos veinte años (2).

Por otra parte, ya tenemos la capacidad de simular en ordenadores la forma en la que las secuencias de pares de bases dieron lugar a secuencias de aminoácidos que a su vez se plegaron en proteínas tridimensionales a partir de las cuales se construye la biología. Asimismo, la complejidad de las proteínas para las cuales podemos realizar simulaciones sobre su manera de plegarse, ha aumentado constantemente a medida que los recursos informáticos han continuado creciendo exponencialmente (3). También podemos simular cómo las proteínas interactúan las unas con las otras siguiendo una intrincada danza tridimensional de fuerzas atómicas. Por tanto, nuestro creciente conocimiento de la biología es un factor importante que nos permite descubrir los secretos inteligentes que la evolución nos ha otorgado y los usos que se les puede dar a estos paradigmas inspirados en la biología a la hora de crear tecnologías cada vez más inteligentes.

En estos momentos, miles de científicos e ingenieros están llevando a cabo un gran proyecto. Su objetivo es comprender el mejor ejemplo que tenemos de un proceso inteligente: el cerebro humano. Así, podría decirse que se trata del mayor esfuerzo en la historia de la civilización hombre-máquina. En *La Singularidad está cerca* defiendo que una consecuencia de la ley de los rendimientos acelerados es que seguramente no existan otras especies inteligentes. Un resumen de este argumento es el siguiente: dado el relativamente corto espacio de tiempo que dista entre el momento en el que una civilización crea tecnología primitiva y el dominio tecnológico que le permite trascender los límites de su planeta (4) (téngase en cuenta que en 1850 la manera más rápida de enviar información a nivel nacional era el Pony Express), si existieran otras especies inteligentes ya las habríamos detectado. Desde esta perspectiva, la aplicación de la ingeniería inversa al cerebro humano puede ser considerada como el proyecto más importante del universo.

El objetivo del proyecto es comprender de forma precisa cómo funciona el cerebro y luego utilizar el desvelamiento de dichos métodos para comprendernos mejor a nosotros mismos, así como para reparar el cerebro cuando sea necesario y (lo que es lo más importante para este libro) para crear máquinas cada vez más inteligentes. Téngase en cuenta que precisamente a lo que se dedica la ingeniería es a amplificar fenómenos naturales. A modo de ejemplo, considérese el fenómeno tan sutil que describe el principio de Bernoulli, que afirma que la presión del aire sobre una superficie curva en movimiento es ligeramente menor que sobre una superficie plana en movimiento. La explicación matemática de cómo el principio de Bernoulli hace que las alas se eleven es algo en lo que todavía los científicos no se han puesto de acuerdo, y sin embargo la ingeniería ha hecho suyo este delicado conocimiento, ha estudiado sus capacidades y ha dado lugar ni más ni menos que al mundo de la

aviación.

Este libro presenta la tesis defendida por lo que yo llamo la teoría de la mente basada en el reconocimiento de patrones (PRTM)*, la cual (según mi opinión) describe el algoritmo básico del neocórtex (la región del cerebro responsable de la percepción, la memoria y el pensamiento crítico). En los siguientes capítulos describo cómo recientes investigaciones en neurociencia, así como nuestros propios experimentos mentales, nos llevan a la irrefutable conclusión de que la utilización de este método abarca toda la extensión del neocórtex y de que este hace un uso sistemático de aquel. Asimismo, la combinación de la LOAR con la PRTM implica que seremos capaces de desarrollar diseños con base en estos principios que nos permitirán aumentar enormemente las capacidades de nuestra propia inteligencia.

De hecho, este proceso ya está muy avanzado. Hay cientos de tareas y actividades que antes eran el dominio exclusivo de la inteligencia humana y que ahora son llevadas a cabo por ordenadores, los cuales suelen ser más precisos que los humanos y pueden trabajar a una escala mucho mayor. Cada vez que usted envía un email o hace una llamada por teléfono móvil, algoritmos inteligentes optimizan el enrutado de la información. Lo mismo ocurre con la realización de electrocardiogramas que son interpretados por ordenador con una calidad que rivaliza con la de los doctores, y este también es el caso en la obtención de imágenes de células sanguíneas. Algoritmos inteligentes detectan automáticamente el fraude con tarjetas de crédito, pilotan y aterrizan aviones, guían sistemas armamentísticos inteligentes, ayudan a diseñar productos mediante diseños asistidos por ordenador, controlan los niveles de los inventarios *just-in-time*, ensamblan productos en fábricas robóticas y practican juegos tales como el ajedrez o incluso el sutil go a nivel de grandes maestros.

Millones de personas presenciaron cómo el ordenador de IBM llamado Watson jugó a *Jeopardy!* en lenguaje natural y obtuvo una mejor puntuación que la de los dos mejores jugadores del mundo juntos. A esto hay que añadir que Watson no solo leyó y "comprendió" el sutil lenguaje utilizado en el cuestionario de *Jeopardy!* (que incluye cosas como juegos de palabras y metáforas), sino que consiguió el conocimiento necesario como para encontrar las respuestas a partir de la comprensión por sí mismo de cientos de millones de páginas pertenecientes a documentos escritos en lenguaje natural que incluyen Wikipedia y otras enciclopedias. Literalmente, necesitó dominar cada área del esfuerzo intelectual humano, incluyendo la historia, la ciencia, la literatura, las artes, la cultura, etc. Actualmente IBM está trabajando con Nuance Speech Technologies (la antigua Kurzweil Computer Products, mi primera empresa) en una nueva versión de Watson que leerá literatura médica (básicamente todas las publicaciones y blogs médicos más importantes) para convertirse en un maestro del diagnóstico y de la consulta

médica. Para ello utilizará las tecnologías de Nuance en la comprensión del lenguaje médico. Algunos analistas han sostenido que Watson realmente no "entiende" los cuestionarios de *Jeopardy!* ni las enciclopedias que ha leído porque se limita a realizar un "análisis estadístico". Un punto fundamental que explicaré en este libro es que las técnicas matemáticas que se han desarrollado en el campo de la inteligencia artificial (tales como las que se usan en Watson y el asistente del iPhone llamado Siri) son matemáticamente muy similares a los métodos que la biología desarrolló bajo la forma del neocórtex. Si entender el lenguaje y otros fenómenos por medio de análisis estadísticos no se considera como verdadero entendimiento, entonces los humanos tampoco poseen entendimiento.

Muy pronto, la capacidad de Watson para dominar inteligentemente conocimientos mediante documentos en lenguaje natural va a ser adoptada por un buscador que usted tiene muy cerca. La gente ya está hablando con sus teléfonos en lenguaje natural, por ejemplo mediante Siri (a cuyo desarrollo también contribuyó Nuance). Así, no falta mucho para que, a medida que utilicen métodos similares a los de Watson y a medida que el propio Watson siga siendo mejorado, estos asistentes en lenguaje natural se vuelvan cada vez más inteligentes.

Los coches autopilotados de Google han cubierto 200 000 millas en las tumultuosas ciudades y pueblos de California (una cifra que sin duda será mucho mayor cuando este libro se encuentre en estanterías tanto reales como virtuales). En el mundo de hoy existen muchos otros ejemplos de inteligencia artificial, y muchos más se divisan en el horizonte.

Como ejemplos adicionales de la LOAR cabe decir que tanto la resolución espacial del escaneo del cerebro y la cantidad de datos que se están reuniendo sobre el cerebro se están doblando anualmente. También se está demostrando que estos datos se pueden convertir en modelos y simulaciones funcionales de las regiones del cerebro. Ya hemos logrado aplicar la ingeniería inversa a funciones fundamentales del córtex auditivo, que es donde procesamos información procedente de sonidos, así como del córtex visual, que es donde procesamos información procedente de nuestro sentido de la vista, y del cerebelo, que es donde realizamos parte de nuestra adopción de habilidades (como por ejemplo, coger al vuelo una pelota).

Lo más ambicioso del proyecto para comprender, modelizar y simular el cerebro humano es aplicar la ingeniería inversa al neocórtex cerebral, que es donde realizamos nuestro jerárquico pensamiento recursivo. El córtex cerebral, que acapara el 80% del cerebro humano, está compuesto de una estructura muy repetitiva, lo cual permite que los humanos creen arbitrariamente ideas dotadas de estructuras complejas.

A partir de la teoría de la mente basada en el reconocimiento de patrones, describo un modelo que explica cómo el cerebro humano adquiere su capacidad crítica mediante una estructura muy inteligente diseñada por la evolución biológica.

Todavía hay detalles de este mecanismo cortical que no comprendemos completamente, pero sabemos lo suficiente sobre las funciones que necesita realizar como para diseñar algoritmos que satisfagan el mismo objetivo. Asimismo, a medida que empezamos a comprender el neocórtex estamos logrando aumentar enormemente su alcance, al igual que la aviación ha aumentado el alcance del principio de Bernoulli. Por tanto, es posible que el principio de funcionamiento del neocórtex sea la idea más importante del mundo, ya que es capaz de representar todo el conocimiento y habilidades humanas, así como de crear nuevos conocimientos. Después de todo, el neocórtex es el responsable de todas las novelas, de todas las canciones, de todos los cuadros, de todos los descubrimientos científicos y de todas las variopintas creaciones del pensamiento humano.

El campo de la neurociencia está muy necesitado de una teoría que englobe las extremadamente dispares y extensas observaciones que diariamente están siendo publicadas. Además, una teoría unificada es una necesidad fundamental de todas las áreas importantes de la ciencia. En el capítulo 1 describiré cómo dos soñadores unificaron la biología y la física, campos que previamente parecían irremediablemente desordenados y diversos, y luego analizaré cómo una teoría de estas características puede ser aplicada al cerebro.

Hoy día solemos encontrarnos con grandes elogios a la complejidad del cerebro humano. Si buscamos citas sobre este tema en Google, se nos muestran 30 millones de enlaces. (Es imposible traducir esto en el verdadero número de citas, ya que algunos sitios web a los que se muestran enlaces pueden contener varias citas o a veces pueden no contener ninguna). El propio James D. Watson escribió en 1992 que "el cerebro es la última y más importante frontera biológica, lo más complejo que hasta ahora hemos encontrado en nuestro universo". Luego explica que él cree que "contiene cientos de miles de millones de células interconectadas a través de billones de conexiones. Así, el cerebro deja atónita a la mente" (5).

Estoy de acuerdo con Watson en cuanto al sentimiento de que el cerebro es la frontera biológica más importante, pero si pudiéramos identificar fácilmente patrones inteligibles (y reproducibles) en sus células y conexiones (especialmente en las que son masivamente redundantes), entonces el hecho de que contenga muchos miles de millones de células y billones de conexiones no significaría necesariamente que su método primario fuera complejo.

Pensemos sobre lo que significa que algo sea complejo. Nos podemos preguntar ¿es complejo un bosque? La respuesta depende de la perspectiva elegida. Se podría señalar que hay muchos miles de árboles en el bosque y que cada uno es diferente. Entonces se podría mencionar que cada árbol posee miles de ramas y que todas son completamente diferentes. Después se podría proceder a describir los retorcidos caprichos de una sola rama. La conclusión podría ser que el bosque posee una

complejidad que supera la imaginación más desbocada.

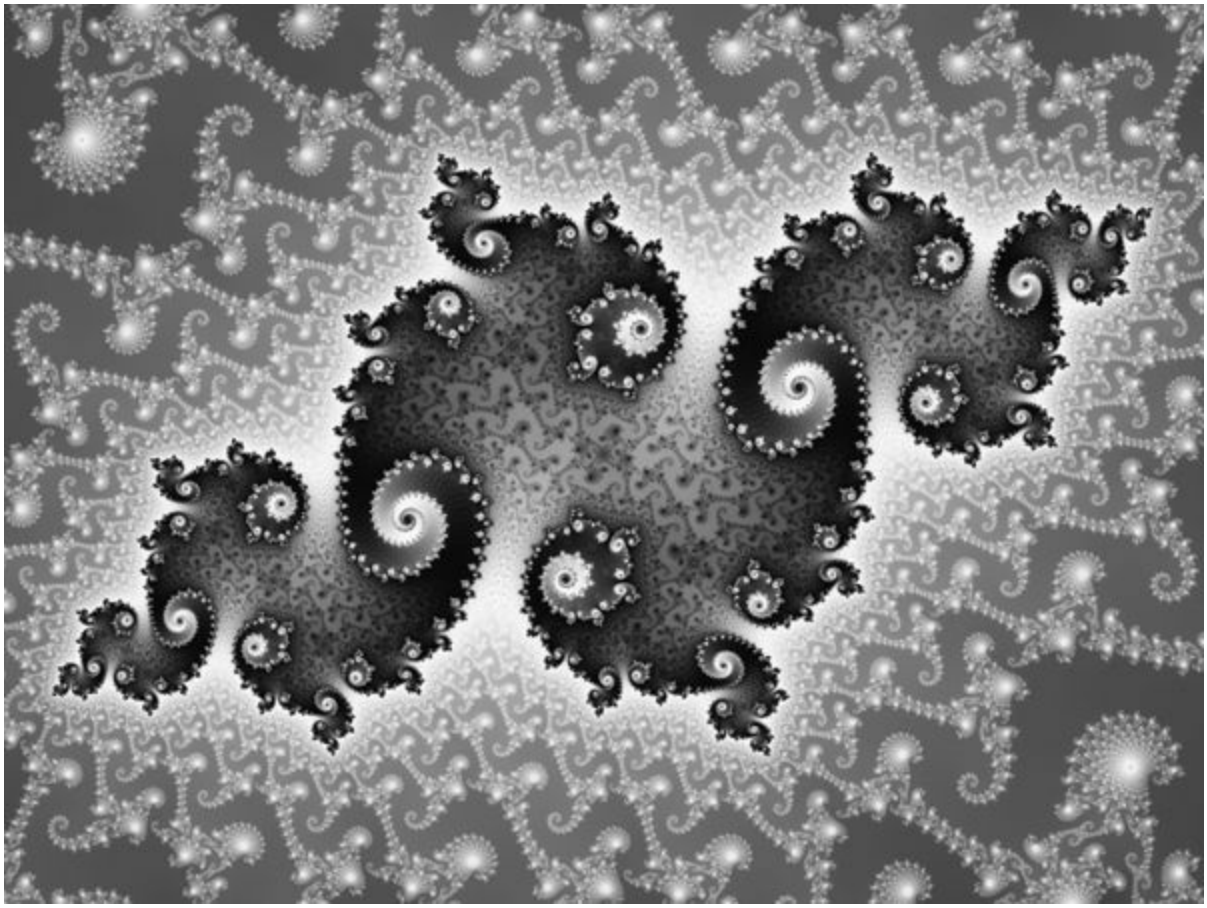
Sin embargo, en una aproximación así los árboles (literalmente) no nos dejarían ver el bosque. Es cierto que hay una gran cantidad de variación fractal encerrada en los árboles y ramas, pero para entender correctamente los principios de un bosque sería mejor empezar por identificar los distintos patrones de redundancia con variación estocástica (es decir, aleatoria) que pudiéramos encontrar. Así, sería justo decir que el concepto de bosque es más sencillo que el concepto de árbol.

Este es el caso del cerebro, el cual posee una enorme redundancia, especialmente en el neocórtex. Tal y como describiré en este libro, sería justo decir que hay más complejidad en una sola neurona que en la estructura general del neocórtex.

Mi objetivo con este libro no es en absoluto añadir una nueva cita a las millones que ya existen y que atestiguan lo complejo que es el cerebro, sino más bien impresionarle a usted con el poder de su simplicidad. Esto lo realizaré describiendo cómo un ingenioso mecanismo básico que se repite cientos de millones de veces y que sirve para reconocer, recordar y predecir un patrón es el responsable de la gran diversidad de nuestro pensamiento. Igual que una asombrosa diversidad de organismos surge de las diferentes combinaciones de valores del código genético que se encuentra en el ADN nuclear y mitocondrial, basada en los valores de los patrones que se encuentran en el interior y entremedias de nuestros reconocedores de patrones, surge una asombrosa variedad de ideas, pensamientos y capacidades. Tal y como dice el neurocientífico del MIT Sebastian Seung, “la identidad no depende de nuestros genes, sino de las conexiones entre nuestras células cerebrales” (6).

Por tanto, tenemos que distinguir entre verdadera complejidad de diseño y complejidad aparente. Consideremos el famoso conjunto de Mandelbrot, cuya imagen es desde hace mucho un símbolo de complejidad. Para apreciar su aparente complejidad es útil aumentar su imagen (a la cual puede usted tener acceso a través del enlace que encontrará en la nota (7)). Se trata de una interminable sucesión de laberintos contenidos en otros laberintos que son siempre diferentes. Sin embargo, el diseño (la fórmula) del conjunto de Mandelbrot no puede ser más simple. Mide seis caracteres: $Z = Z^2 + C$, donde $Z$ es un número “complejo” (un par de números) y $C$ es una constante. No es necesario comprender completamente la función de Mandelbrot para darse cuenta de que es simple. Sin embargo, esta fórmula es aplicada iterativamente y en cada nivel de una jerarquía. Lo mismo pasa con el cerebro. Su repetitiva estructura no es tan simple como la de la fórmula de seis caracteres en el conjunto de Mandelbrot, pero no es ni de lejos tan compleja como sugieren los millones de citas sobre la complejidad del cerebro. Este diseño neocortical se repite una y otra vez en cada nivel de la jerarquía conceptual representada por el neocórtex. Einstein expresó adecuadamente mi objetivo con este

libro cuando dijo que “cualquier necio inteligente puede hacer que las cosas sean más grandes y más complejas […] pero se necesita […] mucho coraje para moverse en la dirección opuesta”.



**Visualización del set de Mandelbrot, una fórmula simple representada iterativamente. Al utilizar el zoom, las imágenes cambian constantemente de forma aparentemente compleja.**

Hasta el momento me he referido al cerebro, ¿pero qué pasa con la mente? Por ejemplo, ¿cómo consigue ser consciente un neocórtex que resuelve problemas? Y ahora que mencionamos esto, ¿cuántas mentes conscientes tenemos en nuestro cerebro? Existen evidencias que sugieren que puede haber más de una.

Otra pregunta pertinente sobre el cerebro es ¿qué es el libre albedrío? ¿poseemos tal cosa? Hay experimentos que parecen demostrar que empezamos a poner en práctica nuestras decisiones antes incluso de que las hayamos tomado. ¿Significa eso que el libre albedrío es una ilusión?

Y por último, ¿qué atributos de nuestro cerebro son los responsables de formar nuestra identidad? ¿soy la misma persona que era hace seis meses? Está claro que no soy exactamente el mismo que entonces, pero ¿tengo la misma identidad?

En relación con estas antiquísimas cuestiones, veremos lo que propone la teoría de la mente basada en el reconocimiento de patrones.

* Libro publicado por Lola Books en español.
** Siglas en inglés correspondientes a “Law Of Accelerated Returns”.

* Siglas en inglés de “pattern recognition theory of mind”.

CAPÍTULO UNO

# Experimentos mentales históricos

La teoría de Darwin de la selección natural nació muy tarde en la historia del pensamiento.

¿Se debió este retraso a que se oponía a la verdad revelada, a que se trataba de un tema absolutamente nuevo en la historia de la ciencia, a que era aplicable solo a los seres vivos o a que se centraba solo en los objetivos y en las causas finales sin postular un origen de la creación? Yo creo que no. Simplemente, Darwin descubrió el papel jugado por la selección, una forma de causalidad muy diferente de los mecanismos de acción-reacción* utilizados por la ciencia hasta ese momento. El origen de una enorme variedad de seres vivos pasó a poder ser explicado mediante la contribución hecha por aquellas nuevas características (seguramente de origen aleatorio) que les habían permitido sobrevivir. En la física o en la biología, muy poco o nada hacía presagiar que la selección fuera un principio causal.

—B. F. SKINNER

En último término, aparte de la integridad de la propia mente nada es sagrado.

—RALPH WALDO EMERSON

## *Una metáfora tomada de la geología*

A principios del siglo XIX, los geólogos se hicieron una pregunta fundamental. Por todo el planeta había grandes cavernas y cañones tales como el Gran Cañón del Colorado en los EE.UU. y el desfiladero de Vikos en Grecia (el cañón más profundo encontrado hasta la fecha). ¿Cómo surgieron estas majestuosas formaciones?

Siempre había una corriente de agua que parecía haber aprovechado la oportunidad para discurrir a través de estas estructuras naturales, pero antes de la mitad del siglo XIX se consideraba absurdo que estos suaves flujos pudieran ser los creadores de valles y acantilados tan enormes. Sin embargo, el geólogo británico Charles Lyell (1797–1875) sugirió que era el movimiento del agua lo que, grano de arena a grano de arena, había esculpido estas enormes modificaciones geológicas a lo largo de grandísimos periodos de tiempo. Al principio, esta sugerencia fue ridiculizada, pero dos décadas después fue mayoritariamente aceptada.

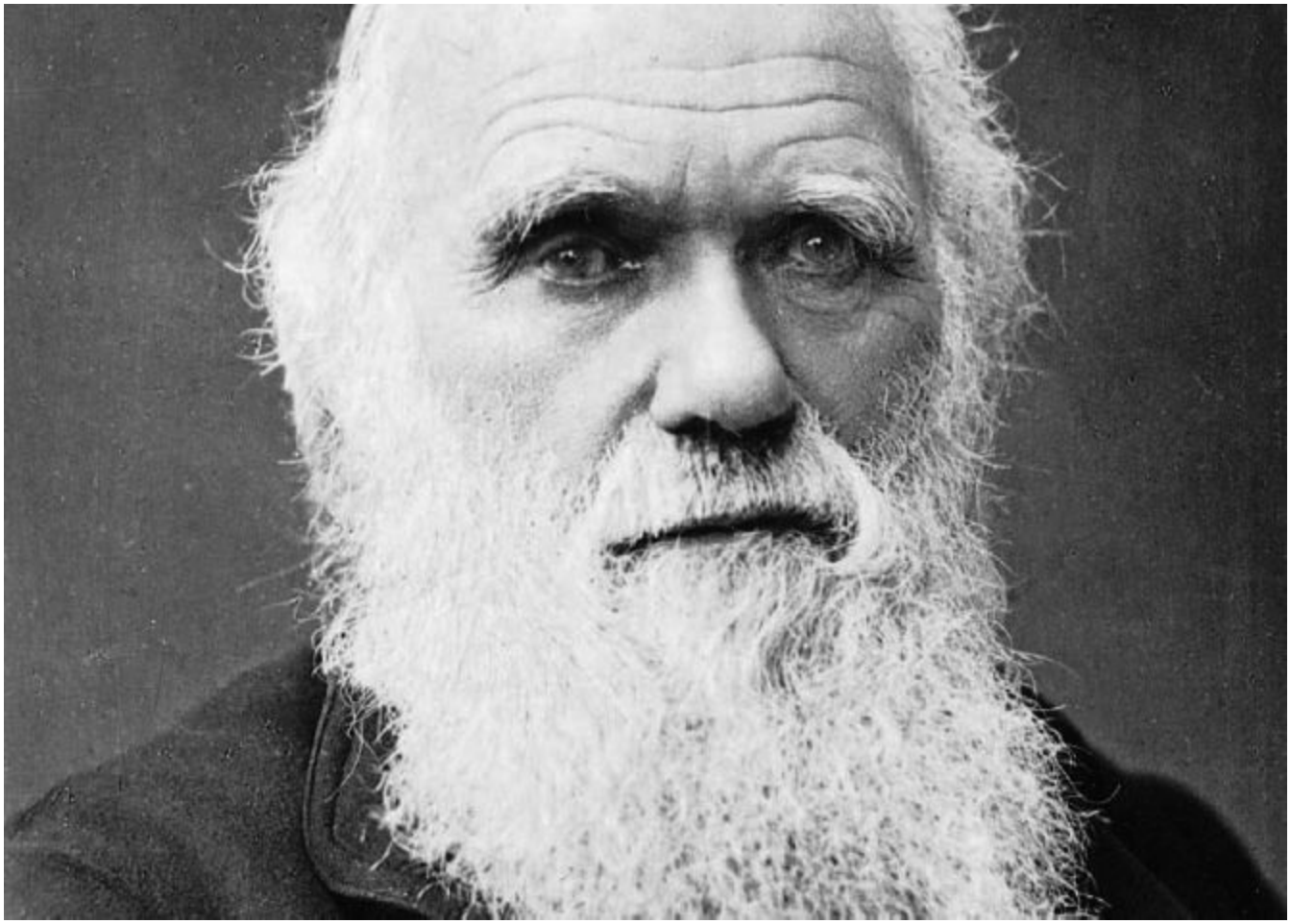
Una persona que observó cuidadosamente la respuesta de la comunidad científica a la tesis radical de Lyell fue el naturalista inglés Charles Darwin (1809–1882). Hay que tener en cuenta la situación de la biología hacia 1850. Se trataba de un campo

infinitamente complejo que se topaba con innumerables especies animales y vegetales, cada una de las cuales presentaba una gran complejidad. Por encima de todo, la mayoría de científicos se resistía a intentar plantear una teoría unificadora de la deslumbrante variedad encontrada en la naturaleza. Esta diversidad servía como testimonio de la gloria de la creación de Dios, y por supuesto de la inteligencia de los científicos que eran capaces de abarcarla.

Darwin abordó el problema concibiendo una teoría general de las especies que era análoga a la tesis de Lyell y con ella explicó los cambios graduales en las características de las especies que se producen a lo largo de muchas generaciones. Después, durante su famoso viaje en el Beagle, combinó esta perspectiva con sus propios experimentos mentales y observaciones. Darwin sostuvo que en cada generación los individuos que sobreviven mejor en su nicho ecológico son los individuos que dan lugar a la siguiente generación.

El 22 de noviembre de 1859 se publicó su libro *El Origen de las Especies*, en el que dejaba clara su deuda con Lyell:

> Me doy perfecta cuenta de que a esta doctrina de la selección natural, considerada por las imaginarias instancias superiores, se le pueden hacer las mismas objeciones que al principio se erigieron contra las nobles opiniones de Sir Charles Lyell cuando postuló "los cambios actuales de La Tierra como evolución geológica". Sin embargo, es muy raro escuchar hoy que, por ejemplo, las olas que azotan la costa sean consideradas como causas insignificantes cuando se habla de la excavación de gigantescos valles o de la formación de los más amplios acantilados de tierra adentro. La selección natural solo puede actuar mediante la preservación y acumulación de modificaciones heredadas infinitamente pequeñas, cada una de las cuales es aprovechada para la preservación del ser vivo en cuestión. Al igual que de la geología moderna casi han desaparecido las opiniones que defienden que un gran valle pueda ser excavado mediante una sola ola diluvial, la selección natural (si es un principio verdadero) acabará con la creencia en la creación continua de nuevos seres orgánicos o en la modificación repentina y profunda de su estructura (1).



**Charles Darwin, autor de *El Origen de las Especies*, que estableció la idea de evolución biológica.**

Siempre hay múltiples razones por las que las nuevas ideas encuentran oposición, y en el caso de Darwin no es difícil identificarlas. A muchos analistas no les sentó bien que no descendiéramos de Dios, sino de los monos y antes de eso de los gusanos. La implicación de que nuestro perro mascota fuera nuestro primo, al igual que la oruga y la planta por la que se mueve (quizá un primo en millonésimo o milmillonésimo grado, pero pariente al fin y al cabo), fue tomado por muchos como una blasfemia.

Sin embargo, la idea cuajó pronto, ya que dotó de coherencia a lo que anteriormente había sido una plétora de observaciones sin relación aparente. Hacia 1872, para la publicación de la sexta edición de *El Origen de las Especies*, Darwin añadió este pasaje: "A modo de crónica de un estado de cosas pasado, he mantenido en los parágrafos anteriores […] varias frases que implican que los naturalistas creen que cada especie se creó por separado. He sido muy censurado por haberme expresado así. Sin embargo, esta manera de pensar era la creencia generalizada cuando este trabajo se presentó por primera vez […]. Ahora las cosas son completamente diferentes y casi todos los naturalistas admiten el gran principio de la evolución" (2).

La idea unificadora de Darwin se acentuó durante el siguiente siglo. Así, en 1869, solo una década después de la primera publicación de *El Origen de las*

*Especies*, el médico suizo Friedrich Miescher (1844–1895) descubrió una sustancia llamada "nucleína" en el núcleo celular que resultó ser el ADN (3). En 1927, el biólogo ruso Nikolai Koltsov (1872–1940) describió lo que él llamaba una "molécula hereditaria gigante" y dijo que estaba compuesta de "dos cadenas especularmente simétricas que se replican de manera semiconservativa usándose la una a la otra como molde". También muchos condenaron su descubrimiento. Los comunistas lo consideraban propaganda fascista, y su repentina e inesperada muerte se ha atribuido a la policía secreta de la Unión Soviética (4). En 1953, casi un siglo después de la publicación del seminal libro de Darwin, el biólogo norteamericano James D. Watson (nacido en 1928) y el biólogo inglés Francis Crick (1916–2004) ofrecieron la primera definición precisa de la estructura del ADN. La describieron como una doble hélice de dos largas moléculas entrelazadas (5). Es preciso reseñar que su descubrimiento se basó en lo que se conoce como "fotografía 51", que fue tomada por su colega Rosalind Franklin usando cristalografía de rayos X. Esta fue la primera representación que mostró la doble hélice. Por eso, dadas las consecuencias que tuvo la imagen de Franklin, se ha sugerido que esta debería haber compartido el Premio Nobel de Watson y Crick (6).

A partir de la descripción de una molécula que podía codificar el programa de la biología, se asentó sobre seguro una teoría unificadora de la biología que proporcionaba una sencilla y elegante base para todo tipo de vida. Así, un organismo puede madurar hasta convertirse en una brizna de césped o en un ser humano dependiendo solamente de los valores que tomen los pares de bases que componen las cadenas de ADN en el núcleo de la célula y, en menor grado, la mitocondria. No obstante, esta perspectiva no acababa con la encantadora diversidad de la naturaleza, sino que nos hacía comprender que su extraordinaria diversidad surge a partir de una gran variedad de estructuras que pueden ser codificadas por esta molécula universal.

## *A lomos de un haz de luz*

A principios del siglo XX, el mundo de la física cambió totalmente gracias a otra serie de experimentos mentales. En 1879, un ingeniero alemán y un ama de casa tuvieron un niño. Este no empezó a hablar hasta que cumplió los tres. Además, se sabe que a los nueve tuvo problemas en el colegio y que a los dieciséis fantaseaba con galopar a lomos de un rayo de luna.

Este joven estaba al corriente del experimento que en 1803 hiciera el matemático inglés Thomas Young (1773–1829) y que demostró que la luz se compone de ondas. En aquel tiempo, la conclusión fue que la luz debía de viajar a través de algún tipo de medio (después de todo, las olas del océano viajaban a través del agua y las

ondas del sonido viajaban a través del aire y de otros materiales). A este medio por el que viajaba la luz los científicos lo llamaron "éter". Nuestro joven también conocía el experimento llevado a cabo en 1887 por los científicos Albert Michelson (1852–1931) y Edward Morley (1838–1923) que intentó confirmar la existencia del éter. La analogía en la que se basaba el experimento era un viaje en una barca de remos que se desplazaba hacia arriba y hacia abajo por el curso de un río. Si se rema a una velocidad constante, la velocidad de la barca medida desde la orilla será mayor si se rema a favor de la corriente que si se rema contracorriente. Además, Michelson y Morley asumían que la luz viajaría a través del éter a velocidad constante (es decir, a la velocidad de la luz).

A partir de esto, su razonamiento les llevó a pensar que la velocidad de la luz del sol cuando La Tierra se desplaza por su órbita en dirección al sol (medida desde nuestro punto de vista terrestre) y su aparente velocidad cuando La Tierra se desplaza en dirección contraria al sol debería ser diferente. La diferencia tendría ser igual al doble de la velocidad de La Tierra. Esto confirmaría la existencia del éter. Sin embargo, lo que descubrieron fue que no había ninguna diferencia en la velocidad de la luz del sol con respecto a La Tierra, independientemente del lugar de la órbita en el que se encontrara esta.

Así, su descubrimiento rebatió la idea del "éter", pero entonces ¿qué era lo que pasaba en realidad? Durante casi dos décadas esto continuó siendo un misterio.

Al igual que parece inmóvil un tren que viaja a nuestro lado y a nuestra misma velocidad, cuando el adolescente alemán se imaginaba cabalgando junto a un haz de luz, creía que debería ver las ondas de la luz inmóviles. Sin embargo, se daba cuenta de que esto era imposible, ya que se supone que la velocidad de la luz es constante independientemente del movimiento de uno mismo. Esto le llevó a imaginar que galopaba junto a un haz de luz pero a una velocidad un poco más baja. ¿Qué pasaría al viajar al 90% de la velocidad de la luz? Si los haces de luz son como trenes, argüía, entonces debería ver el haz de luz adelantándole a una velocidad igual al 10% de la velocidad de la luz. De hecho, eso sería lo que verían los observadores desde la Tierra. Sin embargo, sabemos que la velocidad de la luz es constante, tal y como demostraba el experimento de Michelson-Morley, y por lo tanto esto significaba necesariamente que vería el haz de luz alejándose de él a la velocidad de la luz. Esto parecía una contradicción, ¿cómo podría ser posible algo así?

El chaval alemán, que por cierto se llamaba Albert (1879–1955), tuvo clara la respuesta al cumplir los veintiséis, momento en el que al gran Einstein le se le hizo evidente que *para él el tiempo en sí tendría que discurrir más despacio*. Explica su razonamiento en un trabajo publicado en 1905 (7). Si observadores terrestres pudieran ver el reloj del joven, verían que funciona diez veces más despacio. De hecho, cuando volviera a La Tierra su reloj mostraría un intervalo de tiempo de solo el 10% (dejando a un lado por el momento la aceleración y la desaceleración).

Desde su perspectiva, sin embargo, su reloj habría funcionado normalmente y el haz de luz de su lado se habría desplazado a la velocidad de la luz. El 10% de disminución en la velocidad del tiempo en sí (comparado con los relojes de La Tierra) explica perfectamente las aparentes discrepancias en cuanto a la perspectiva.

En último término, la disminución en el paso del tiempo llegaría a ser cero si la marcha alcanzase la velocidad de la luz, lo cual significaba que era imposible cabalgar a la misma velocidad que la luz. No obstante y pese a esto, teóricamente no resultaba imposible moverse más rápido que el haz de luz, por lo que si se pudiera adelantar al haz de luz el tiempo iría hacia atrás.

Muchas de las primeras críticas tildaron esto de absurdo. ¿Cómo podría ralentizarse el tiempo basándose solamente en la velocidad de desplazamiento de uno mismo? Así, desde que dieciocho años antes se realizara el experimento Michelson-Morley, el resto de pensadores habían sido incapaces de llegar a la conclusión que al gran Einstein le parecía tan obvia. Todos los que habían sopesado este problema durante la última parte del siglo XIX "se habían caído del caballo" en términos del seguimiento de las implicaciones de un principio, ya que optaron por adherirse a sus preconcebidas nociones sobre cómo debe funcionar la realidad. (Seguramente mi metáfora debería decir que "se cayeron del haz de luz").

El segundo experimento mental de Einstein fue imaginarse que él y su hermano volaban por el espacio. Les separan 186 000 millas. Einstein quiere moverse más deprisa, pero al mismo tiempo desea que la distancia entre él y su hermano siga siendo la misma. Por eso, cada vez que quiere acelerar, le hace a su hermano una señal luminosa. Dado que él sabe que la señal luminosa tardará un segundo en llegar a su hermano, aguarda un segundo después de haber hecho la señal para empezar a acelerar. Su hermano siempre acelera inmediatamente después de haber recibido la señal. De esta manera, ambos hermanos aceleran exactamente al mismo tiempo y por tanto la distancia entre los dos se mantiene constante.

Sin embargo, consideremos ahora lo que veríamos desde la Tierra. Si los hermanos se estuvieran alejando de nosotros con Albert a la cabeza, parecería que la señal tardara menos de un segundo en alcanzar al hermano, ya que este está desplazándose hacia la luz. Asimismo, podríamos observar cómo el reloj del hermano de Albert disminuye su velocidad (si se desplazara hacia nosotros, veríamos cómo la velocidad del reloj aumenta). Por estas dos razones, veríamos cómo los dos hermanos se acercan cada vez más y al final colisionan. Sin embargo, desde la perspectiva de los dos hermanos, ambos permanecen separados a una distancia de 186 000 millas.

¿Cómo puede ser esto posible? La respuesta es que (*obviamente*) las distancias se contraen paralelamente al movimiento (no perpendicularmente). Esto significa que (dando por sentado que vuelan con la cabeza hacia delante) los hermanos Einstein se

hacen cada vez más pequeños a medida que se desplazan más deprisa. En un principio, es probable que con esta extraña conclusión Einstein perdiera más seguidores que con la conclusión relativa al paso del tiempo.

Ese mismo año, Einstein sopesó la relación entre materia y energía mediante otro experimento mental más. En la década de 1850, el físico escocés James Clerk Maxwell había demostrado que las partículas de la luz llamadas fotones no tenían masa, pero que sin embargo tenían momento. Cuando era pequeño tuve un aparato llamado radiómetro de Crookes (8) que consistía en una bombilla de cristal hermética que contenía un vacío parcial y un conjunto de cuatro veletas que rotaban alrededor de un huso. Las veletas eran blancas por un lado y negras por el otro. El lado blanco de cada veleta reflejaba la luz y el negro absorbía la luz. (Por eso es más fresco llevar una camiseta blanca que una negra en un día de calor). Cuando una luz alumbraba el aparato, las veletas rotaban según los lados oscuros se alejaban de la luz. Esta es una demostración fehaciente de que los fotones portan el momento suficiente como para hacer que las veletas del radiómetro se muevan (9).

La cuestión a la que se enfrentó Einstein es que el momento se define en función de la masa: el momento es igual a la masa multiplicada por la velocidad. Dado que una locomotora desplazándose a 30 millas por hora tiene mucho más momento que, pongamos por caso, un insecto desplazándose a la misma velocidad, ¿cómo puede entonces ser posible que una partícula de masa igual a cero tenga momento positivo?

El experimento mental de Einstein consistía en una caja flotando en el espacio. Un fotón es emitido en el interior de la caja desde el lado izquierdo hacia el derecho. Necesariamente, el momento del sistema ha de conservarse, de manera que la caja tendría que retroceder hacia la izquierda cuando el fotón fuera emitido. Después de un cierto tiempo, el fotón colisiona contra el lado derecho de la caja, transfiriendo así su momento de vuelta a la caja. De nuevo, el momento total del sistema se conserva, de manera que la caja deja de moverse.

Por ahora, todo bien. Sin embargo, téngase en cuenta el punto de vista privilegiado del Sr. Einstein, que está observando la caja desde el exterior. No observa ninguna influencia externa sobre la caja, ya que ninguna partícula (con o sin masa) impacta sobre ella y nada sale de la caja. No obstante, el Sr. Einstein, según el escenario descrito anteriormente, ve cómo la caja se mueve temporalmente hacia la izquierda y luego se detiene. De acuerdo con nuestro análisis, cada fotón debería mover la caja permanentemente hacia la izquierda. Además, dado que no ha habido efectos extraños sobre la caja o desde el interior de ella, su centro de masas debe permanecer en el mismo lugar. Sin embargo, el fotón en el interior de la caja, que se mueve de izquierda a derecha, no puede cambiar el centro de masas porque no tiene masa, ¿o sí que puede? La conclusión de Einstein fue que, dado que obviamente el fotón posee energía y momento, también debe de tener una masa equivalente. La energía de un fotón en movimiento es por completo equivalente a una masa en

movimiento. Teniendo en cuenta que el centro de masas del sistema tiene que permanecer estacionario durante el movimiento del fotón, podemos calcular la equivalencia. Al hacer cuadrar matemáticamente este razonamiento, Einstein demostró que la masa y la energía son equivalentes y se relacionan entre sí según una constante simple. Sin embargo, aquí había trampa, ya que la constante puede que fuera simple, pero resultó ser enorme: el cuadrado de la velocidad de la luz (más o menos $1{,}7 \cdot 10^{17}$ metros$^2$ por segundo$^2$, es decir, 17 seguido de 16 ceros). Así es cómo se obtiene la famosa ecuación de Einstein $E=mc^2$ (10). Por consiguiente, una onza (28 gramos) de masa equivale a 600 000 toneladas de trinitrotolueno (TNT). La carta que envió Einstein el 2 de agosto de 1939 informaba al Presidente Roosevelt del potencial para crear una bomba atómica encerrado en esta fórmula, lo cual dio lugar a la era nuclear (11).

**Un radiómetro de Crookes— la veleta de cuatro alas rota cuando la luz la ilumina.**

Se podría pensar que esto ya tenía que estar claro anteriormente, ya que los investigadores ya se habían dado cuenta de que con el tiempo la masa de las sustancias radioactivas disminuía como resultado de la radiación. Sin embargo, se daba por sentado que las sustancias radioactivas contenían un cierto tipo de combustible de alta energía que iban consumiendo progresivamente. Esta suposición

no es del todo errónea, lo que pasa es que el combustible “consumido” es, simplemente y llanamente, masa.

Hay varias razones por las cuales he empezado este libro con los experimentos mentales de Darwin e Einstein. Primero, muestran el extraordinario poder del cerebro humano. Sin otro equipamiento que un lápiz y un papel para dibujar los esquemas de estos simples experimentos mentales y para anotar las razonablemente sencillas ecuaciones que surgen de ellos, Einstein fue capaz de derrocar una manera de comprender el mundo físico que se había mantenido durante dos siglos. Además, influenció enormemente el curso de la historia (incluyendo la segunda guerra mundial) y nos abocó a la era nuclear.

Es verdad que Einstein se basó en unos pocos descubrimientos del siglo XIX, pero estos experimentos tampoco hacían uso de un equipamiento sofisticado. También es cierto que la posterior demostración experimental de las teorías de Einstein sí que ha hecho uso de tecnologías avanzadas, y que si estas no hubieran sido desarrolladas no hubiéramos obtenido la demostración que tenemos hoy en día de que las ideas de Einstein son ciertas y relevantes. Sin embargo, dichos factores no le restan importancia al hecho de que estos famosos experimentos mentales revelan el poder del pensamiento humano en su máxima expresión.

Está mayoritariamente asumido que Einstein fue el científico más importante del siglo XX (Darwin podría ser un buen candidato para el mismo honor en lo que respecta al siglo XIX). Empero, los razonamientos matemáticos que subyacen de sus teorías no son muy complicados. Los experimentos mentales en sí fueron sencillos. Así, nos podríamos preguntar en qué aspecto se le puede considerar a Einstein particularmente inteligente. Por eso más adelante discutiremos exactamente lo que estaba haciendo con su cerebro cuando se le ocurrieron estas teorías y dónde reside dicha cualidad.

Por otra parte, esta historia también demuestra las limitaciones del pensamiento humano. Einstein fue capaz de cabalgar sobre su haz de luz sin caerse (aunque llegó a la conclusión de que era imposible cabalgar sobre un haz de luz), pero ¿cuántos miles de analistas y pensadores fueron absolutamente incapaces de reflexionar adecuadamente sobre estos ejercicios tan sorprendentemente sencillos? Un error extendido es la dificultad que la mayor parte de la gente tiene a la hora de descartar y trascender las ideas y opiniones de sus contemporáneos. También existen otras deficiencias que examinaremos con más detalle una vez que hayamos examinado cómo funciona el neocórtex.

## *Un modelo unificado del neocórtex*

La razón más importante por la cual hago referencia a los que puede que sean los experimentos mentales más famosos de la historia es que pueden ser utilizados a modo de introducción sobre lo que respecta al cerebro. Tal y como se verá, por medio de unos cuantos experimentos de nuestra cosecha podemos llegar sorprendentemente lejos a la hora de explicar cómo funciona la inteligencia humana. Por eso, teniendo en cuenta la materia que nos ocupa, los experimentos mentales deberían servirnos como ejemplo de estrategia a seguir.

Si los despreocupados pensamientos de un muchacho y un equipamiento de tan solo un lápiz y un papel fueron suficientes como para revolucionar nuestra manera de entender la física, entonces deberíamos poder hacer un progreso razonable en lo que respecta a un fenómeno con el que estamos mucho más familiarizados. Después de todo, experimentamos nuestro pensamiento en cada momento de nuestra vida despierta así como durante los sueños.

Después de que gracias a este proceso de autorreflexión hayamos construido un modelo sobre el funcionamiento del pensamiento, examinaremos hasta qué punto podemos confirmarlo por medio de las observaciones más recientes hechas sobre cerebros reales y haciendo uso de procedimientos de vanguardia cuyo objetivo es recrear estos procesos en máquinas.

* “Push-pull mechanisms” en el original.

CAPÍTULO DOS

# Experimentos mentales sobre el pensamiento

Casi nunca pienso mediante palabras. Un pensamiento aparece y a lo mejor intento expresarlo en palabras posteriormente.

—ALBERT EINSTEIN

El cerebro es una masa de tres libras, que se puede sujetar en una mano, capaz de concebir un universo que mide cien mil millones de años luz.

—MARIAN DIAMOND

Lo que parece asombroso es que un mero objeto de tres libras hecho de los mismos átomos que constituyen todo lo que existe bajo el sol sea capaz de dirigir todo lo que los humanos hemos hecho: volar a la Luna y conseguir setenta *home runs*, escribir *Hamlet*, construir el Taj Mahal e incluso descubrir los secretos del propio cerebro.

—JOEL HAVEMANN

Empecé a pensar sobre el pensamiento allá por el año 1960, el mismo año que descubrí los ordenadores. A día de hoy, sería difícil encontrar a una persona de doce años que no utilice un ordenador, pero en aquel entonces en mi ciudad natal (Nueva York) solo había un puñado de estas máquinas. Por supuesto, estos primitivos dispositivos no cabían en la mano, el primero al que tuve acceso ocupaba una habitación grande. A principios de la década de 1960, llevé a cabo ciertas programaciones en un IBM 1620 con el objetivo de realizar análisis de varianza (una prueba estadística) sobre datos recogidos mediante el estudio de un programa destinado a la educación infantil temprana, un precursor de *Head Start**. Lo cierto es que en esta tarea había involucrada una importante cantidad de dramatismo, ya que el destino de esta iniciativa educativa a nivel nacional dependía de nuestro trabajo. Los algoritmos y los datos analizados eran lo suficientemente complejos como para que no pudiéramos anticipar las respuestas que daría el ordenador. Las respuestas, obviamente, dependían de los datos, pero no eran predecibles, ya que la diferencia entre estar *determinado* y ser *predecible* es importante (volveré sobre ella).

Me acuerdo lo excitante que era ver cómo parpadeaban las luces del panel frontal justo antes de que el algoritmo terminara sus deliberaciones. Parecía como si

el ordenador estuviera profundamente concentrado. Cuando alguien pasaba por allí con ganas de pasar al siguiente conjunto de respuestas, yo me limitaba a señalar las tenues luces parpadeantes y decía: "está pensando". Se trataba al mismo tiempo de un chiste y de una frase en serio, ya que *sí* que parecía estar sopesando las respuestas, por lo que los componentes del grupo de trabajo empezaron a atribuir cierta personalidad a la máquina. Puede que se tratara de una analogía antropomórfica, pero consiguió que empezara a plantearme seriamente la relación entre pensamiento y ordenador.

Para evaluar el grado de similitud entre mi propio cerebro y los programas de ordenador con los que estaba familiarizado, empecé a pensar sobre lo que mi cerebro tenía que realizar a la hora de procesar información. Esta investigación la he prolongado durante cincuenta años. Lo que a continuación voy a señalar en relación a nuestro conocimiento actual sobre el funcionamiento del cerebro parecerá diferir mucho del concepto habitual que se tiene sobre los ordenadores. Sin embargo, en lo fundamental, el cerebro almacena y procesa información, y debido a la universalidad de la computación, un concepto sobre el que volveré a referirme, existen más paralelismos entre el cerebro y los ordenadores de lo que pudiera parecer.

Cada vez que hago o pienso algo, ya sea lavarme los dientes, pasear por la cocina, plantearme un problema empresarial, practicar sobre un teclado musical o formular una nueva idea, reflexiono sobre cómo he sido capaz de conseguirlo. Y pienso todavía más sobre todas las cosas que no soy capaz de hacer, dado que las limitaciones del pensamiento humano nos proporcionan un conjunto de indicios que es igualmente importante. Es posible que pensar tanto sobre el pensamiento me haga ir más despacio. No obstante, mantengo la esperanza de que dicho ejercicio de introspección me permita refinar mis métodos mentales.

Para tomar mayor conciencia sobre cómo funciona nuestro cerebro, consideremos una serie de experimentos mentales.

Pruebe usted lo siguiente: *recite el alfabeto*.

Seguramente lo recordará desde la infancia y lo pueda recitar fácilmente.

Muy bien, pues ahora intente lo siguiente: *recite el alfabeto al revés*.

A menos que usted haya estudiado el alfabeto en dicho orden, es muy posible que considere esto como algo imposible. Si alguien ha pasado una considerable cantidad de tiempo en un aula de educación primaria en la que el alfabeto estuviera a la vista, entonces podría apelar a su memoria visual y a partir de ahí leerlo hacia atrás. Pero incluso esto es difícil, ya que en realidad no recordamos imágenes completas. Aun así, recitar el alfabeto hacia atrás debería ser una tarea sencilla, ya que se trata de exactamente la misma información que cuando se recita hacia adelante. Sin embargo, por lo general no lo logramos.

¿Recuerda usted su número de la Seguridad Social? Si así fuera, ¿podría usted

recitarlo hacia atrás sin tener que escribirlo antes? ¿Y podría hacer lo mismo con la canción de guardería *Mary Had a Little Lamb?* Esto se logra con los ordenadores de forma rutinaria. Sin embargo, nosotros no lo logramos a no ser que aprendamos la secuencia del revés, como si se tratara de una nueva serie. Esto nos indica algo importante sobre cómo se organiza la memoria humana.

Ciertamente, esta tarea la podemos realizar fácilmente si primero escribimos la secuencia y luego la leemos al revés. Al hacerlo, estamos utilizando un tipo de tecnología (el lenguaje escrito) para compensar una de las limitaciones de nuestro desprovisto pensamiento, aunque se trate de una herramienta muy rudimentaria. (Se trata de nuestra segunda invención después del lenguaje hablado). Por eso inventamos herramientas, para compensar nuestras carencias.

Esto sugiere que **nuestros recuerdos son secuenciales y que están sujetos a un orden. Además, son accesibles en el orden en el que son recordados y no podemos revertir directamente la secuencia de un recuerdo.**

También tenemos dificultades a la hora de activar un recuerdo en mitad de una secuencia. Si aprendo a tocar una pieza para piano, normalmente no puedo empezar a tocarla a partir de cualquier punto. Sí que puedo hacerlo en determinados puntos, ya que mi recuerdo secuencial de la pieza está organizado en segmentos. Sin embargo, si intento empezar en mitad de un segmento, tengo que acudir a la lectura de las notas hasta para que mi recuerdo secuencial surta efecto.

Ahora intente usted lo siguiente: *recuerde un paseo que usted haya dado durante el día de ayer o recientemente. ¿Qué es lo que recuerda?*

Este experimento mental funciona mejor si el paseo es muy reciente, de ayer o de hoy. (También vale cambiar el paseo por una conducción o por cualquier actividad en la que se haya desplazado sobre el terreno).

Es probable que no recuerde mucho sobre esta experiencia. ¿Quién fue la primera persona con la que se encontró (sin incluir solamente a la gente que conoce)? ¿Vio algún roble? ¿un buzón? ¿qué fue lo que vio al doblar la primera esquina? Si pasó por alguna tienda, ¿qué es lo que había en el segundo escaparate? Quizás pueda usted reconstruir la respuesta a alguna de estas cuestiones partiendo del recuerdo de ciertas pistas, pero es probable que recuerde relativamente pocos detalles, aunque se trate de una experiencia muy reciente.

Si usted pasea habitualmente, remóntese al primer paseo que dio el mes pasado (o al primer viaje a la oficina del mes pasado, si es que va en transporte público). Probablemente, no pueda recordar en absoluto este paseo o viaje en concreto, pero si pudiera hacerlo sin duda recordaría todavía menos detalles sobre él que sobre el paseo de hoy. Más adelante discutiré la cuestión de la consciencia y defenderé que tendemos a equiparar consciencia con recuerdos de acontecimientos. Así, la principal razón por la que creemos no estar conscientes cuando estamos anestesiados

es que no podemos recordar nada de ese periodo (aunque se producen intrigantes y perturbadoras excepciones). Pero entonces, en lo que respecta al paseo de esta mañana, ¿es posible que no estuviera consciente durante la mayor parte del tiempo? Se trata de una pregunta razonable, dado que no recuerdo casi nada de lo que vi o de lo que pensé.

Sin embargo, sí que hay algunas cosas que recuerdo de mi paseo de esta mañana. Me acuerdo que pensé sobre este libro, aunque no podría decir exactamente qué es lo que pensé. También me acuerdo de pasar al lado de una mujer que empujaba un carrito de bebé. Me acuerdo de que la mujer era atractiva y de que el bebé era muy mono. Me acuerdo de dos pensamientos con los que asocié esta experiencia. Primero pensé: *este bebé es adorable, igual que mi nieto*, y luego: *¿qué estará percibiendo el bebé en su campo visual?* No recuerdo ni lo que llevaban puesto, ni el color su pelo (mi mujer diría que esto es típico en mi). Aunque no puedo describir nada específico sobre su apariencia, sí que tengo algún sentimiento inefable sobre el aspecto de la madre y creo que sería capaz de reconocer su foto entre varias. De manera que, aunque que debe de haber algo con respecto a su aspecto que he retenido en mi memoria, si pienso en la mujer, el carrito y el bebé no soy capaz de visualizarlos. En mi mente no existe ninguna fotografía o vídeo del evento. Por tanto, es difícil describir exactamente *lo que hay* en mi mente en lo que respecta a esta experiencia.

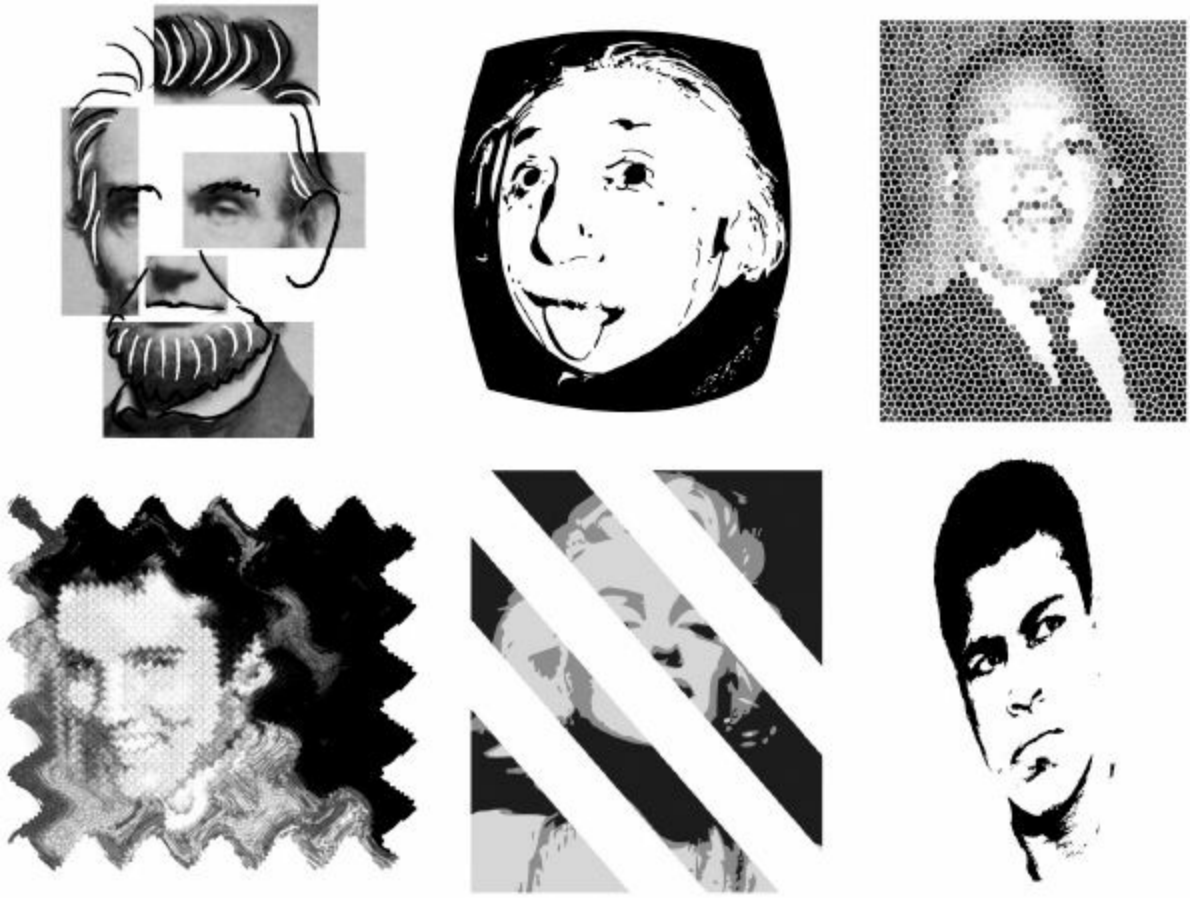
También recuerdo que, durante un paseo que di hace unas pocas semanas, me crucé con una mujer diferente pero que también empujaba un carrito de bebé. En este caso, no creo que ni siquiera fuera capaz de reconocer la fotografía de dicha mujer. Ese recuerdo debe ser ahora mucho más tenue que poco después del paseo.

*A continuación, piense en personas con las que solo se ha encontrado una o dos veces*. ¿Puede visualizarlas claramente? Si usted es un artista visual, puede que haya adquirido esta habilidad observacional, pero por lo general somos incapaces de visualizar con la suficiente claridad a personas con las que nos hemos cruzado de forma ocasional como para poder dibujarlas o describirlas. Sin embargo, solemos poder reconocerlas en una foto sin mucha dificultad.

Esto nos sugiere que **no hay grabaciones de imágenes, ni de vídeos, ni de sonidos almacenadas en nuestro cerebro. Nuestros recuerdos están almacenados a modo de secuencias de patrones. Así, las memorias que no se rememoran se debilitan con el tiempo.** Cuando los retratistas de la policía interrogan a la víctima de un delito, no preguntan "¿qué aspecto tenían las cejas del delincuente?" Sino que muestran una serie de imágenes de cejas y piden a la víctima que elija una. El conjunto de cejas correcto desencadenará el reconocimiento del mismo patrón que está almacenado en la memoria de la víctima.

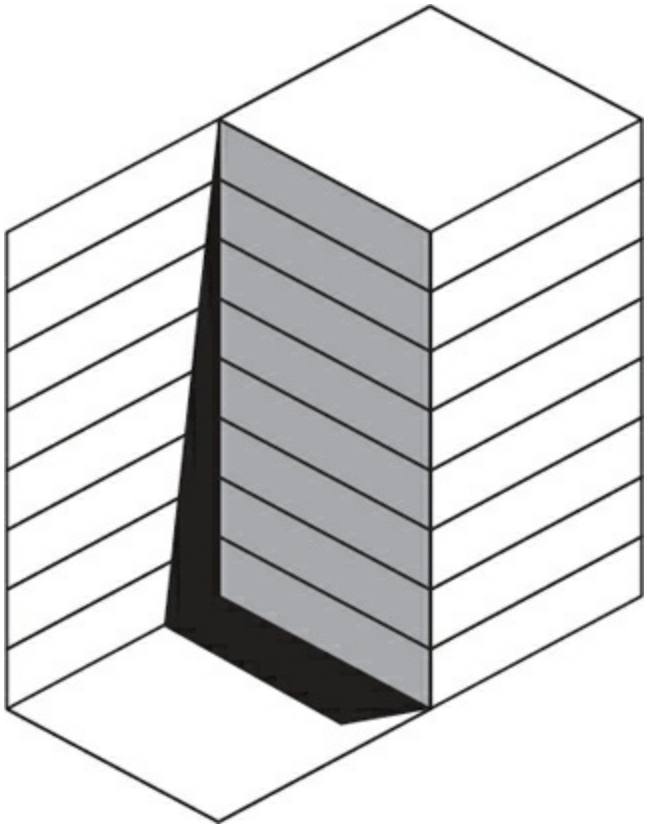
Consideremos ahora caras que usted conozca bien. *¿Podría usted reconocer a*

*estas personas?*



Sin duda, aunque estén parcialmente cubiertas o distorsionadas, usted será capaz de reconocer a estas personas, ya que le son familiares. Esto representa una capacidad fundamental de la percepción humana: **incluso si solo percibimos una parte de él (mediante la vista, el oído o el tacto), podemos reconocer un patrón, aunque este contenga alteraciones. Aparentemente, nuestra capacidad de reconocimiento es capaz de detectar características invariables de un patrón (aquellas características que perduran más allá de las variaciones en el mundo real).** Las distorsiones en la apariencia que se dan en una caricatura o en ciertas formas de arte como por ejemplo el impresionismo, enfatizan los patrones que reconocemos en una imagen (ya se trate de una persona o de un objeto) aunque otros detalles cambien. De hecho, el mundo del arte le lleva ventaja al mundo de la ciencia a la hora de apreciar la capacidad del sistema perceptivo humano. También utilizamos esta misma estrategia cuando reconocemos una melodía a partir de unas pocas notas.

*A continuación, deténgase en esta imagen:*



La imagen es ambigua— la esquina que viene indicada por la región en negro podría ser una esquina interior o una esquina exterior. La primera vez es probable que la perciba de una manera y luego de otra, aunque si hace un esfuerzo puede cambiar su percepción y alternar las interpretaciones. Sin embargo, una vez que su mente se centre en una de las interpretaciones, puede ser difícil ver la otra perspectiva. (Esto también es así en lo que se refiere a las perspectivas intelectuales). De hecho, la interpretación que hace su cerebro de la imagen influye en la manera en la que usted la percibe. Cuando la esquina parece ser interior, su cerebro interpreta la región gris como si fuera una sombra, de manera que no parece ser tan oscura como cuando interpreta que la esquina es exterior.

Por tanto, podemos decir que **la experiencia consciente de nuestras percepciones cambia según las interpretaciones que hagamos.**

*Tenga en cuenta que vemos lo que esperamos _ _ _*

Estoy convencido de que usted ha sido capaz de completar la frase anterior.

Si hubiera escrito la última palabra, usted solo habría necesitado un breve vistazo para confirmar que era la palabra que usted esperaba.

Esto implica que **constantemente estamos prediciendo el futuro y haciendo hipótesis sobre lo que vamos a experimentar. Esta expectativa influye sobre lo que de hecho percibimos.** Tanto es así que la predicción del futuro es la principal razón por la que poseemos un cerebro.

*Considere una experiencia que todos tenemos habitualmente: un recuerdo sobre algo que pasó hace años que inexplicablemente surge en su cabeza.*

Suele tratarse de un recuerdo sobre una persona o acontecimiento sobre el que no ha pensado desde hace mucho. Es evidente que algo ha provocado dicho recuerdo. Además, es posible que el tren de pensamientos que lo provocó se haga patente y que usted sea capaz de articularlo. Sin embargo, otras veces puede que usted sea consciente de la secuencia de pensamientos que le condujo al recuerdo, pero que en cambio le resulte difícil expresarlo. Y no solo eso. A menudo el desencadenante se desvanece rápidamente, de manera que el recuerdo parece haber surgido de la nada. Yo suelo experimentar estos recuerdos aleatorios mientras estoy haciendo tareas rutinarias tales como lavarme los dientes. A veces puede que sea consciente de la conexión (la pasta de dientes cayéndose del cepillo puede que me recuerde a la pintura que se cayó del pincel durante una clase de dibujo que di en la universidad). Sin embargo, a veces solo tengo una noción vaga de la conexión, o incluso ninguna en absoluto.

Un fenómeno relacionado con esto que todo el mundo experimenta frecuentemente se produce al intentar recordar un nombre o una palabra. En este caso, el procedimiento que utilizamos es intentar acordarnos de los desencadenantes que pueden liberar un recuerdo. (Por ejemplo, *¿quién hizo de la reina Padmé en* La venganza de los sith? *Vamos a ver, se trata de la misma actriz que protagonizó una oscura película reciente cuyo tema era el baile, la película era* Cisne negro, *ah, sí, se trata de Natalie Portman*). Otras veces adoptamos nuestras propias reglas nemotécnicas que nos ayudan a recordar. (Por ejemplo, *siempre está delgada, no rolliza, ah, sí, Portman, Natalie Portman*)*. Algunos de nuestros recuerdos son lo suficientemente fuertes como para que vayamos directamente desde una pregunta (por ejemplo, ¿quién *hizo de reina Padmé?*) hasta la respuesta. Sin embargo, a menudo necesitamos pasar por una serie de desencadenantes hasta encontrar el adecuado. Esto se parece mucho a dar con el enlace web correcto. De hecho, los recuerdos pueden perderse igual que una página web a la que no se accede a través de ningún enlace (o por lo menos ningún enlace que podamos encontrar).

*Cuando realice tareas rutinarias como ponerse una camisa, obsérvese a sí mismo realizándolas y reflexione hasta qué punto sigue usted la misma secuencia cada vez que las realiza*. Basándome en la observación sobre mí mismo (como ya he dicho, intento observarme constantemente), es probable que siga en gran medida los mismos pasos cada vez que realice una tarea rutinaria en particular, aunque se pueden añadir componentes adicionales. Por ejemplo, la mayoría de mis camisas no necesitan de gemelos, pero una sí que los necesita, cosa que implica una nueva serie de tareas.

En mi mente, la lista de los pasos a seguir está organizada según jerarquías. Así, sigo un procedimiento rutinario antes de irme a dormir. El primer paso es lavarme los dientes. Sin embargo, está acción se divide a su vez en una serie de pequeños

pasos, el primero de los cuales es poner pasta de dientes en el cepillo. A su vez, este paso está compuesto de pasos todavía más pequeños, tales como encontrar la pasta de dientes, quitarle la tapa, etc. El paso de encontrar la pasta de dientes también tiene pasos, el primero de los cuales es abrir el armarito del baño. Asimismo, este paso requiere de otros pasos, el primero de los cuales es agarrar la parte exterior de la puerta del armarito. De hecho, esta concatenación continúa hasta llegar a movimientos muy sutiles, de manera que, literalmente, existen miles de pequeñas acciones que constituyen mi rutina nocturna. Aunque pueda tener dificultades a la hora de recordar los detalles de un paseo que di hace tan solo unas cuantas horas, no tengo ninguna dificultad en acordarme de todos estos pequeños pasos antes de meterme en la cama, tanto es así que soy capaz de pensar en otras cosas mientras realizo estos procedimientos.

Es importante señalar que esta lista no está almacenada como una lista compuesta de miles de pasos, **en vez de eso, cada uno de los procedimientos de la rutina es recordado como una jerarquía compuesta de actividades concatenadas.**

**Este mismo tipo de jerarquización tiene que ver con nuestra capacidad de reconocer objetos y situaciones.** Reconocemos las caras de las personas a las que conocemos bien y también nos damos cuenta de que dichas caras contienen ojos, nariz, boca, etc.: una jerarquía de patrones que utilizamos tanto en nuestras percepciones como en nuestras acciones. Además, el uso de jerarquías nos permite reutilizar patrones. Por ejemplo, no tenemos que volver a aprender el concepto de nariz o de boca cada vez que nos encontramos con una nueva cara.

En el siguiente capítulo reuniremos los resultados de estos experimentos mentales en una teoría para explicar la manera en la que debe funcionar el neocórtex. Mi tesis es que estos experimentos revelan atributos esenciales de nuestro pensamiento, y que estos atributos son uniformes y se aplican tanto para encontrar la pasta de dientes como para escribir un poema.

* Programa del Departamento de Salud y Servicios Sociales de los EE.UU.

* Aquí hay que entender el juego de palabras en inglés. El autor utiliza la palabra *portly* (rolliza) y eso le lleva a acordarse del apellido Portman.

CAPÍTULO TRES

# Un modelo del neocórtex: la teoría de la mente basada en el reconocimiento de patrones

El cerebro es un tejido. Es un tejido complicado e intricadamente hurdido que no es comparable con nada conocido en el universo. Sin embargo, al tratarse de un tejido, está compuesto de células. En concreto, se trata de células altamente especializadas, pero que funcionan según las leyes que rigen cualquier otra célula. Sus señales eléctricas y químicas pueden ser detectadas, registradas e interpretadas, y su composición química puede ser identificada. Asimismo, las conexiones que constituyen su intrincada red de fibras nerviosas pueden ser cartografiadas. En resumen, el cerebro puede ser estudiado igual que puede serlo un riñón.

—DAVID H. HUBEL, NEUROCIENTÍFICO

Supongamos que tuviéramos una máquina cuya estructura produjera pensamiento, sensaciones y percepciones. Imaginemos ahora esta máquina aumentada de tamaño, pero preservando las mismas proporciones, de manera que se pudiera entrar en ella como si fuera un molino. Digamos además que se nos permite visitarla por dentro. ¿Qué es lo que veríamos allí? Solamente partes que se empujan y mueven las unas a las otras, pero nada que pudiera explicar la percerción.

—GOTTFRIED WILHELM LEIBNIZ

## *Una jerarquía de patrones*

He repetido los sencillos experimentos y observaciones descritos en el capítulo anterior miles de veces en infinidad de contextos. Al igual que los sencillos experimentos con el tiempo, el espacio y la masa llevados a cabo durante el siglo XIX influyeron decisivamente en las reflexiones del joven maestro Einstein sobre cómo funcionaba el universo, las conclusiones que he sacado de estas observaciones no pueden sino condicionar mi forma de explicacar la manera en la que creo que el cerebro debe de funcionar. Asimismo, en la siguiente exposición también incluiré algunas observaciones muy básicas procedentes de la neurociencia intentando evitar los muchos detalles que todavía están por dilucidar.

Primero, permítaseme explicar por qué esta sección trata específicamente del neocórtex (palabra que en latín significa “anillo nuevo”). Sabemos que el neocórtex

es responsable de nuestra capacidad para manejarnos con patrones de información y para hacerlo de forma jerárquica. Animales sin neocórtex (fundamentalemente los no mamíferos) son claramente incapaces de comprender las jerarquías (1). El entendimiento y uso de la innata naturaleza jerárquica de la realidad es un atributo único de los mamíferos que resulta de la posesión en exclusiva de esta estructura cerebral evolutivamente tan reciente. El neocórtex es el responsable de la percepción sensorial, del reconocimiento de todo (desde objetos visuales a conceptos abstractos), del control del movimiento, del razonamiento basado tanto en la orientación espacial como en el pensamiento racional y del lenguaje (sobre todo en lo que se refiere a lo que llamamos "pensamiento").

El neocórtex humano, la capa externa del cerebro, es una estructura fina y fundamentalmente bidimensional con un grosor de unos 2,5 mm (más o menos una décima de pulgada). En los roedores, tiene aproximadamente el tamaño de un sello y es terso. Una innovación evolutiva de los primates es que el suyo acabó por plegarse de modo intrincado sobre el resto del cerebro, dando lugar a profundas crestas, grutas y arrugas que aumentaron su superficie. Debido a este complicado pliegue, el neocórtex constituye la mayor parte del cerebro humano, ya que es el responsable del 80% de su masa. Los *homo sapiens* desarrollaron una amplia frente que permitió un neocórtex todavía más grande. Concretamente, poseemos un lóbulo frontal en el que nos encargamos de los patrones de mayor abstracción, aquellos asociados a conceptos de alto nivel.

Fundamentalmente, esta fina estructura se compone de seis capas, numeradas del I (la capa exterior) hasta el VI. Los axones que emergen de las neuronas de la capa II y III se proyectan hacia otras partes del neocórtex. Los axones (conexiones de salida) pertenecientes a las capas V y VI se conectan fundamentalmente al exterior del neocórtex, sobre todo al tálamo, al tronco del encéfalo y a la médula espinal. Las neuronas de la capa VI reciben conexiones sinápticas (de entrada) procedentes de las neuronas que están fuera del neocórtex, sobre todo en el tálamo. Además, el número de capas varía ligeramente de región a región. La capa IV en el córtex motor es muy fina, ya que en dicha área no se reciben señales procedentes ni del tálamo, ni del tronco del encéfalo, ni de la médula espinal. Por el contrario, en el lóbulo occipital (la parte del neocórtex que suele ser responsable del procesamiento visual) pueden observarse tres subcapas adicionales pertenecientes a la capa IV, ya que el flujo de señales que llegan a esta región, incluyendo señales procedentes del tálamo, es muy considerable.

Una característica fundamental del neocórtex es la extraordinaria uniformidad de su estructura básica. Esto fue observado por primera vez por el neurocientífico norteamericano Vernon Mountcastle (nacido en 1918). En 1957, Mountcastle descubrió la organización columnar del neocórtex. En 1978 realizó una observación tan importante para la neurociencia como lo fue para la física el experimento

refutatorio de Michelson-Morley sobre el éter en 1887. Aquel año, Mountcastle describió la organización tan extraordinariamente invariable del neocórtex y lanzó la hipótesis de que este estaba compuesto de un único mecanismo que se repetía una y otra vez (2). Además, sugirió que la columna cortical era dicha unidad básica. (Las diferencias en la altura de ciertas capas pertenecientes a las regiones a las que se hace referencia más arriba son simplemente diferencias en la cantidad de interconectividad a las que dichas regiones tienen que hacer frente).

Mountcastle lanzó la hipótesis de la existencia de minicolumnas dentro de las columnas, pero esta teoría estuvo rodeada de mucha controversia, ya que no había demarcaciones visibles que separaran dichas estructuras más pequeñas. Sin embargo, la exhaustiva experimentación ha revelado que de hecho existen unidades que se repiten en el interior de las fábricas neuronales de cada columna. Mi opinión es que dicha unidad básica es un reconocedor de patrones, y que esto es lo que constituye el componente fundamental del neocórtex. A diferencia de lo que postula la idea de Mountcastle sobre las minicolumnas, no existe ninguna barrera física específica entre estos reconocedores, ya que un reconocedor y el siguiente se encuentran muy cerca el uno de otro, conformando una estructura entretejida de manera que la columna cortical es simplemente la suma de un gran número de estos reconocedores de patrones. Estos reconocedores son capaces de conectarse entre sí durante el transcurso de la vida, de manera que la intrincada conectividad entre módulos que observamos en el neocórtex no viene especificada de antemano en el código genético, sino que más bien se crea para reflejar los patrones que aprendemos con el tiempo. Más adelante describiré esta tesis con más detalle, pero no obstante sostengo que es así cómo el neocórtex debe de estár organizado.

Antes de adentrarnos más en la estructura del neocórtex, hay que reseñar que es importante modelizar los sistemas al nivel adecuado. Aunque teóricamente la química se basa en la física y aquella podría ser derivada completamente de esta, esto sería en la práctica poco manejable e impracticable, de manera que la química ha establecido sus propias reglas y modelos. De forma similar, deberíamos ser capaces de deducir las leyes de la termodinámica a partir de la física, pero una vez que dejamos de tener un mero conjunto de partículas y tenemos el suficiente número de ellas como para ponerles el nombre de gas, el resolver las ecuaciones de la interacción física de cada partícula se vuelve algo imposible, mientras que las leyes de la termodinámica funcionan bastante bien. Asimismo, la biología tiene sus propias reglas y modelos. Aislada, una célula de la isleta pancreática es enormemente complicada, sobre todo si la modelizamos a nivel molecular. Sin embargo, modelizar la función que realiza el páncreas en términos de la regulación de los niveles de insulina y de las enzimas digestivas es algo considerablemente menos complejo.

Los mismos principios rigen los niveles de modelización y comprensión del cerebro. Una parte ciertamente útil y necesaria de la aplicación de la ingeniería inversa al cerebro es modelizar sus interacciones a nivel molecular, sin embargo el principal objetivo es el refinamiento de nuestro modelo para poder explicar la manera en la que el cerebro procesa información a la hora de producir significado a nivel cognitivo.

El científico norteamericano Herbert A. Simon (1916–2001), considerado cofundador del campo de la inteligencia artificial, escribió elocuentemente sobre la cuestión de la comprensión de los sistemas complejos al nivel de abstracción adecuado. En 1973, al describir un programa de IA que había creado y llamado EPAM (*elementary perceiver and memorizer*), escribió: "supongamos que usted decide que quiere comprender el misterioso programa EPAM que he creado. Yo podría darle dos versiones del mismo. Una sería […] la forma en la que el programa fue escrito, con toda su estructura de rutinas y subrutinas […]. Por otro lado, le podría dar una versión de EPAM en lenguaje máquina después de haberlo traducido por completo (por decirlo de alguna manera, después de haberlo suavizado). […] No creo necesario discutir mucho sobre cuál de estas dos versiones proporcionaría la descripción más escueta, más dotada de significado y más legítima. […] Ni siquiera le ofreceré una tercera posibilidad, […] no darle ninguno de los programas anteriores, sino las ecuaciones electromagnéticas y el conjunto de condiciones establecidas para el conjunto de ecuaciones diferenciales que el ordenador, tomado como sistema físico, tendría que obedecer mientras se comoporte como EPAM. Eso sería el sumun de la simplificación y de la incomprensibilidad" (3).

En el neocórtex humano hay más o menos medio millón de columnas corticales, cada una de las cuales ocupa unos dos milímetros de alto y medio milímetro de ancho, y contiene unas 60 000 neuronas (lo que da un total de unos 30 mil millones de neuronas en el neocórtex). Una estimación aproximada nos dice que cada reconocedor de patrones contenido en una columna cortical posee alrededor de 100 neuronas, de manera que hay un total aproximado de 300 millones de reconocedores de patrones en el neocórtex.

A la hora de considerar cómo funcionan estos reconocedores de patrones, permítaseme empezar diciendo que es difícil saber con precisión por dónde empezar. En el neocórtex todo ocurre simultáneamente, de manera que el proceso no tiene ni principio ni final. Así, a menudo tendré que referirme a fenómenos que no he explicado todavía pero sobre los que tengo pensado volver, de manera que le ruego que acepte estas referencias a apartados posteriores del libro.

La capacidad de los seres humanos para los procesos lógicos es muy limitada, pero sin embargo poseemos una profunda capacidad para reconocer patrones. Así, para razonar lógicamente necesitamos utilizar el neocórtex, que básicamente es un

gran reconocedor de patrones. No es el mecanismo ideal para realizar transformaciones lógicas, pero es la única instalación que tenemos para realizarlas. Por ejemplo, compárese cómo juega al ajedrez un humano con el funcionamiento del típico programa de ajedrez. Deep Blue, la computadora que derrotó a Garry Kasparov, el campeón del mundo humano, era capaz en 1997 de analizar por segundo las implicaciones lógicas de 200 millones posiciones sobre el tablero (entendidas estas como las diferentes secuencias de jugadas y contrajugadas. Nivel que por cierto ya ha sido alcanzado por algunos ordenadores personales).

A Kasparov se le preguntó cuántas posiciones podía analizar por segundo y respondió que menos de una. Entonces, ¿cómo fue capaz de siquiera plantarle batalla a Deep Blue? La respuesta reside en la fuerte capacidad de los humanos a la hora de reconocer patrones. Sin embargo, esta facilidad la tenemos que entrenar, razón por la cual no todo el mundo alcanza la maestría en el ajedrez. Kasparov había aprendido alrededor de 100 000 posiciones sobre el tablero. Se trata de un número certero, ya que hemos llegado a la conclusión de que un humano experto en cualquier campo domina alrededor de 100 000 fragmentos de conocimiento. Shakespeare compuso sus obras mediante el significado de 100 000 palabras (empleó unas 29 000 palabras diferentes, pero la mayoría las utilizaba de diferentes formas). Además, los sistemas expertos en medicina que han sido creados para representar el conocimiento de un médico humano han demostrado que el especialista médico humano llega a manajar, de media, unos 100 000 conceptos pertenecientes a su campo. Por tanto, reconocer un fragmento de conocimiento en medio de este almacén no es algo sencillo, ya que un ítem en particular se muestra ligeramente diferente cada vez que se experimenta.

Armado con sus conocimientos, Kasparov mira el tablero y compara los patrones que ve con las 100 000 posiciones que ha llegado a dominar, y además realiza todas estas 100 000 comparaciones simultáneamente. Sobre este punto sí que hay consenso: todas nuestras neuronas procesan (sopesan patrones) al mismo tiempo. Esto no quiere decir que todas las neuronas *se disparen* simultáneamente (si eso pasara, seguramente nos caeríamos al suelo), pero mientras procesan sopesan la posibilidad de dispararse.

¿Cuántos patrones puede almacenar el neocórtex? Es necesario incluir aquí el fenómeno de la redundancia. Por ejemplo, la cara de un ser querido no está almacenada una vez, sino miles de veces. Algunas de estas repeticiones son básicamente la misma imagen de la cara, aunque la mayoría muestra diferentes perspectivas de la misma, diferente luminosidad, expresiones diferentes, etc. Ninguno de estos patrones repetidos está almacenado como imagen per se (es decir, como una colección bidimensional de píxeles). En vez de eso, están almacenados como listas de características en las que los elementos constituyentes de un patrón también son patrones. Más adelante decribiremos con más precisión el aspecto que toman estas jerarquías tan características y cómo se organizan.

Si tomamos el conocimiento enraizado en un experto como consistente en unos 100 000 “trozos” de conocimiento (es decir, patrones) con una redundancia estimada de alrededor de 100 a 1, esto nos arroja una necesidad de 10 millones de patrones. Este núcleo de conocimientos expertos está construido sobre conocimientos profesionales más generales y amplios, de manera que podemos aumentar el orden de magnitud de los patrones hasta los 30 o 50 millones. Nuestros conocimientos “de sentido común” cotidiano son todavía mayores (la pericia para orientarse en la calle requiere de nuestro neocórtex bastante más que la pericia para orientarse en los libros). Si incluimos esto, nuestra estimación llega a sobrepasar con creces los 100 millones de patrones, contado con un factor de redundancia de más o menos 100. Téngase en cuenta que el factor de redundancia está muy lejos de ser fijo, ya que los patrones muy comunes poseen un factor de redundancia muy superior a mil, mientras que un fenómeno nuevo puede tener un factor de redundancia de menos de 10.

Tal y como expondré más adelante, nuestra forma de actuar y nuestras acciones también dependen de patrones y también están almacenadas en regiones del neocórtex, de manera que mi estimación sobre la capacidad total del neocórtex humano es del orden de unos pocos cientos de millones de patrones. Este cómputo aproximado se corresponde bastante bien con el número de reconocedores de patrones que he estimado anteriormente (unos 300 millones), de manera que es razonable concluir que la función de cada reconocedor neocortical de patrones es la de procesar una iteración correspondiente a un patrón (es decir, una copia entre las múltiples copias redundantes correspondientes a la mayoría de los patrones del neocórtex). Por tanto, nuestra estimación del número de patrones que el cerebro humano es capaz de manejar (incluyendo la redundancia necesaria) y el número de reconocedores de patrones físicos resultan ser del mismo orden de magnitud. Hay que señalar aquí que cuando me refiero a “procesar” un patrón, me estoy refiriendo a todas las cosas que somos capaces de hacer con un patrón: aprenderlo, predecirlo (incluyendo partes de él), reconocerlo e implementarlo (bien reflexionando sobre él o bien mediante un patrón de movimiento físico).

Trescientos millones de procesadores de patrones puede parecer un número muy alto, y de hecho fueron suficientes como para permitir que el *homo sapiens* desarrollara el lengauje verbal y escrito, todas nuestras herramientas y otras creaciones de diversa índole. Estos inventos se han construido los unos sobre los otros, lo cual dio lugar al crecimiento exponencial de la información contenida en nuestra tecnología, tal y como lo describo en mi ley de los rendimientos acelerados. Ninguna otra especie ha sido capaz de esto. Tal y como he expuesto, unas pocas de otras especies, tales como los chimpancés, parecen poseer una rudimentaria capacidad para comprender y formar lenguaje y para para utilizar herramientas primitivas. Después de todo, también poseen un neocórtex, pero, debido a su tamaño

más pequeño, sus capacidades son limitadas, especialmente en lo que se refiere al lóbulo frontal. El tamaño de nuestro neocórtex ha superado un umbral que le ha permitido a nuestra especie construir herramientas cada vez más poderosas, incluyendo herramientas que ya nos permiten comprender nuestra propia inteligencia. En último término, nuestros cerebros, combinados con las tecnologías que han promovido, nos permitirán crear un neocórtex sintético que contendrá muchos más de 300 millones de procesadores de patrones. ¿Por qué no mil millones? ¿Por qué no un billón?

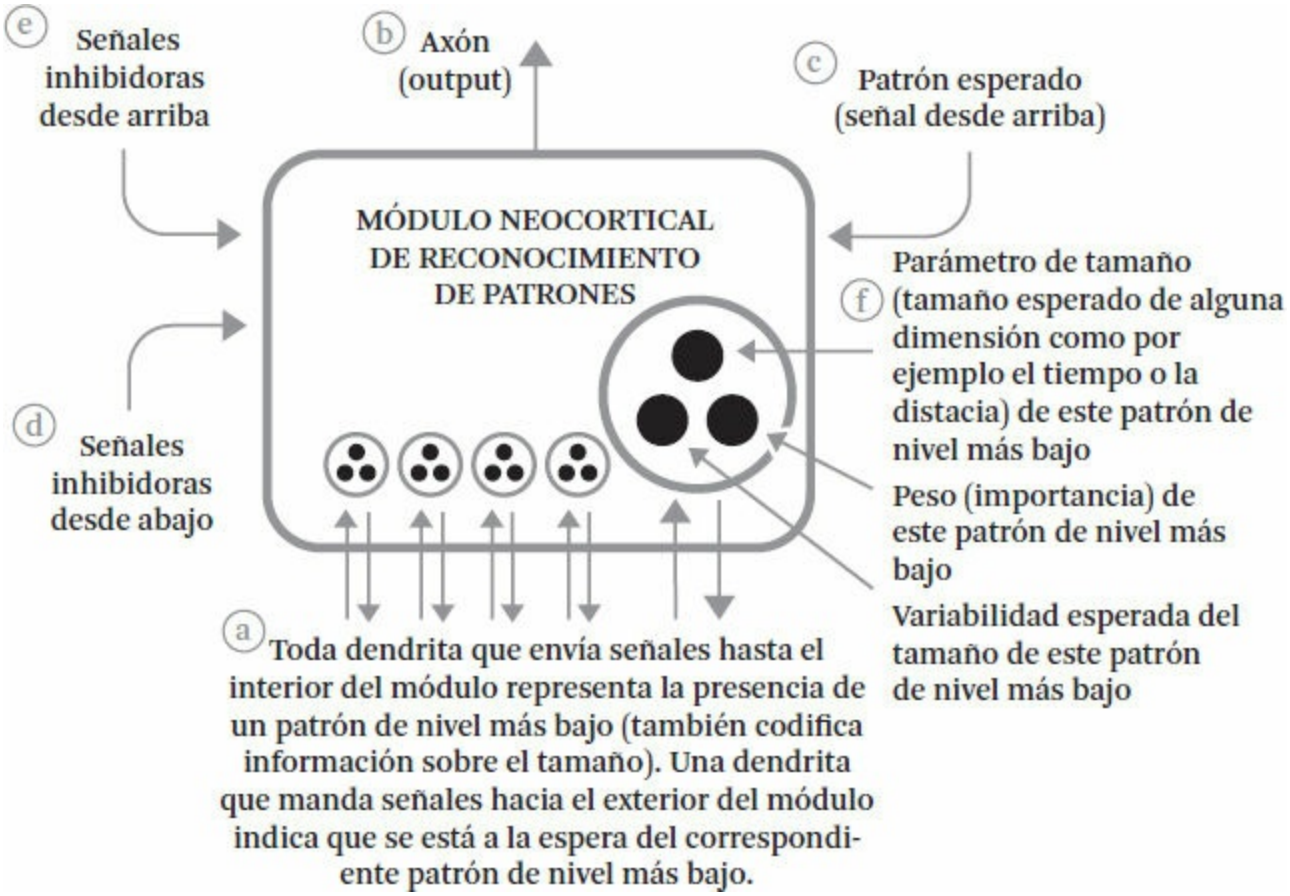
## *La estructura de un patrón*

La teoría de la mente basada en el reconocimiento de patrones que presento aquí se basa en el reconocimiento de patrones por medio de módulos de reconocimiento de patrones en el neocórtex. Estos patrones (y los módulos) se organizan en jerarquías. Más adelante expongo las raízes intelectuales de esta idea, incluyendo mi propia labor en el reconocimiento de patrones jerárquicos en las décadas de 1980 y 1990, así como el modelo del neocórtex llevado a cabo por Jeff Hawkins (nacido en 1957) y por Dileep George (nacido en 1977) a principios de la década de 2 000.

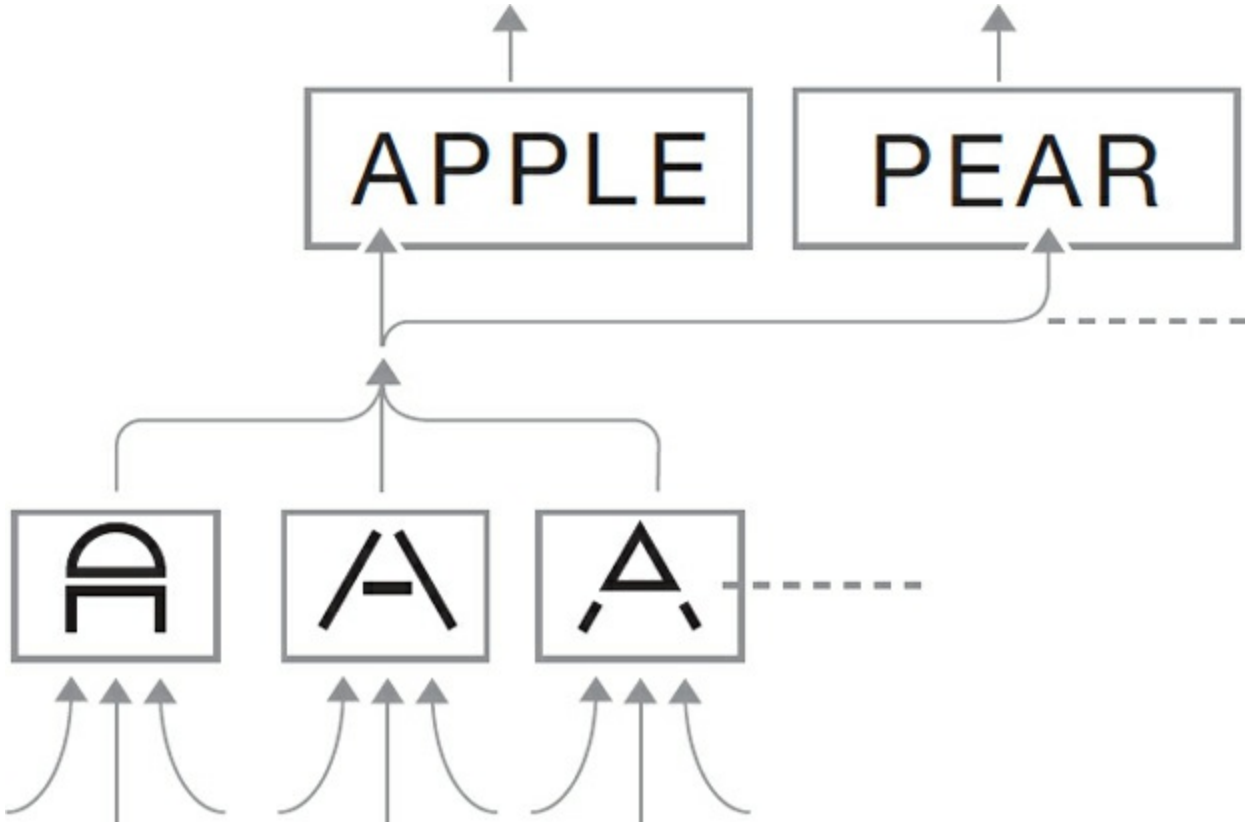
Cada patrón (reconocido por uno de los aproximadamente 300 millones de reconocedores de patrones del neocórtex) se compone de tres partes. La primera parte es el input, que consiste en los patrones de nivel más bajo que componen el patrón principal. Las descripciones de cada uno de estos patrones de bajo nivel no tienen por qué repetirse para cada patrón de nivel más alto a los que hagan referencia. Por ejemplo, muchos de los patrones de las palabras incluirán la letra “A”. Cada uno de estos patrones no necesita repetir la descripción de la letra “A” sino que utilizará la misma descripción. Piense en ello como si se tratara de un puntero web. Existe una página web (es decir, un patrón) para la letra “A” y todas las páginas web (patrones) para las palabras que incluyan la “A” tendrán que enlazar con la página “A” (con el patrón “A”). En vez de enlaces web, el neocórtex utiliza conexiones neuronales reales. Hay un axón que parte del reconocedor del patrón “A” que se conecta con múltiples dendritas, una por cada palabra que utilice la “A”. Tenga también en cuenta el factor de la redundancia, ya que existe más de un reconocedor de patrones para la letra “A”. Por tanto, cualquiera de estos múltiples reconocedores del patrón “A” puede enviar una señal hasta los reconocedores de patrones que incorporan la “A”.

La segunda parte de cada patrón es su nombre. En el ámbito del lenguaje, este patrón de nivel más alto es simplemente la palabra “manzana”. Aunque utilizamos directamente el neocórtex para comprender y procesar cada nivel del lenguaje, la

mayoría de los patrones contenidos en él no son en sí mismos patrones del lenguaje. En el neocórtex el "nombre" de un patrón es simplemente el axón que emerge desde cada procesador de patrones. Así, cuando un axón se dispara es que su patrón correspondiente ha sido reconocido. Por tanto, el disparo del axón es el reconocedor del patrón gritando el nombre del susodicho partrón: "oye, chicos, acabo de ver la palabra escrita 'manzana'".




La tercera y última parte de cada patrón es el conjunto de patrones de más alto nivel del que a su vez él mismo forma parte. Para la letra "A", esto significa todas las palabras que incluyen la "A". De nuevo, esto es similar al caso de los enlaces web. Cada patrón reconocido a un determinado nivel desencadena el siguiente nivel en el que parte de ese patrón de más alto nivel está presente. En el neocórtex, estos enlaces vienen representados por dendritas físicas que fluyen hasta el interior de las neuronas de cada reconocedor de patrones cortical. Tenga en cuenta que cada neurona puede recibir inputs procedentes de múltiples dendritas y producir un solo output en un axón. Sin embargo, el axón puede a su vez mandar señales a múltiples dendritas.




**Tres patrones redundantes (pero en cierto sentido diferentes) de la "A" que sirven para alimentar patrones de nivel más alto que incorporan la "A".***

A modo de ejemplos simples, los sencillos patrones de la siguiente página representan un pequeño subgrupo de los patrones utilizados para crear letras impresas. Fíjese en que cada nivel constituye un patrón. En este caso, las formas son patrones, las letras son patrones y las palabras también lo son. Cada uno de estos patrones posee un conjunto de inputs, un proceso de reconocimiento de patrones (basado en los inputs que tienen lugar en el módulo) y un output (que alimenta el siguiente reconocedor de patrones de nivel más alto).

Conexión entre el sudoeste y el norte:

Conexión entre el sudeste y el norte:

Barra horizontal:

Línea a la izquierda completamente vertical:

Región cóncava que mira hacia el sur:

Línea horizontal en la base:

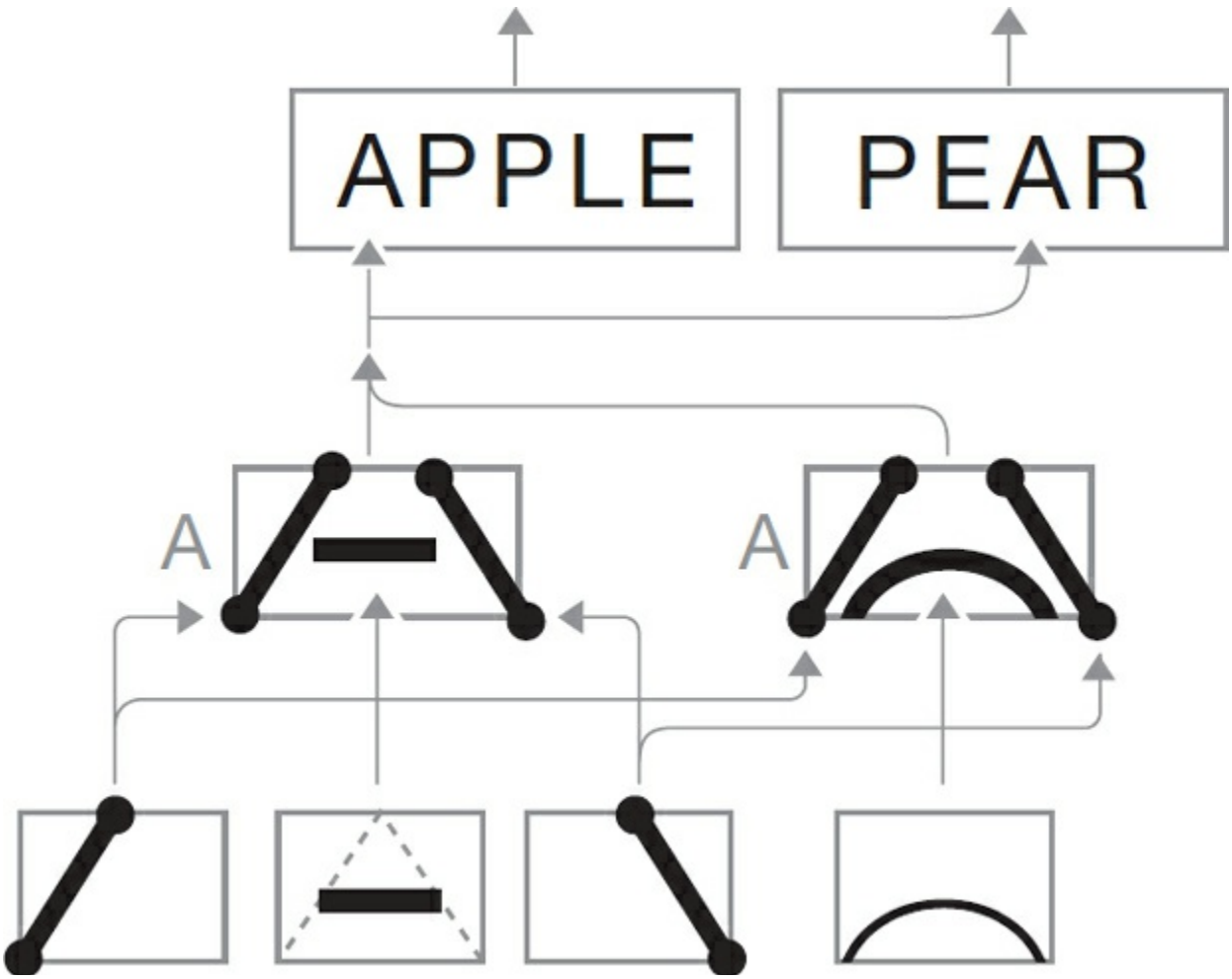
Línea horizontal en la parte de arriba:

Línea horizontal en el medio:

Trayectoria circular que constituye la región superior:

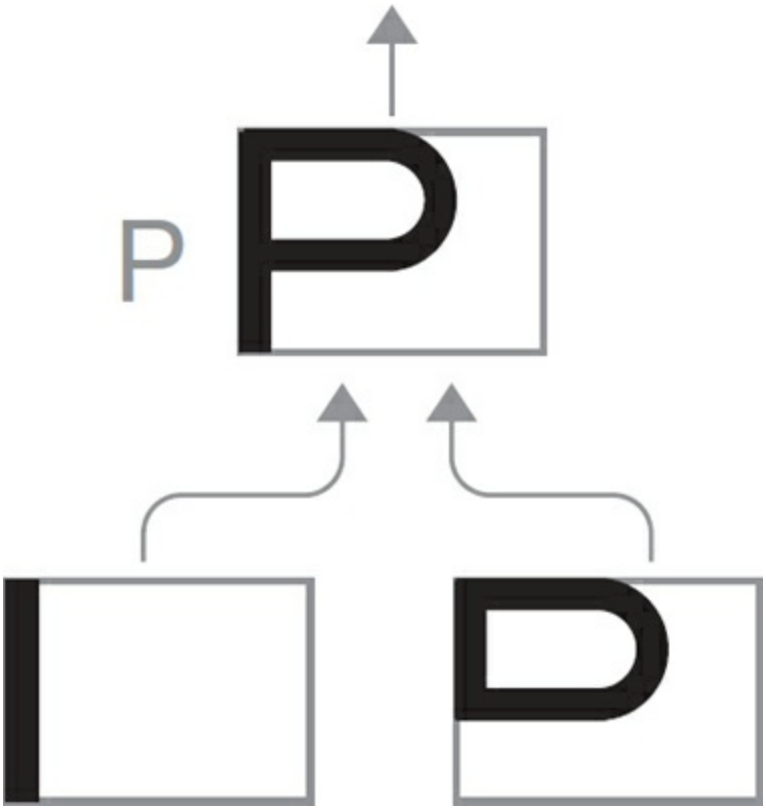
Los patrones anteriores son elementos del siguiente patrón de nivel más alto, que es una categoría llamada letras impresas (formalmente no existe una categoría así en el interior del neocórtex, ya que de hecho no existen las categorías formales).

"A":



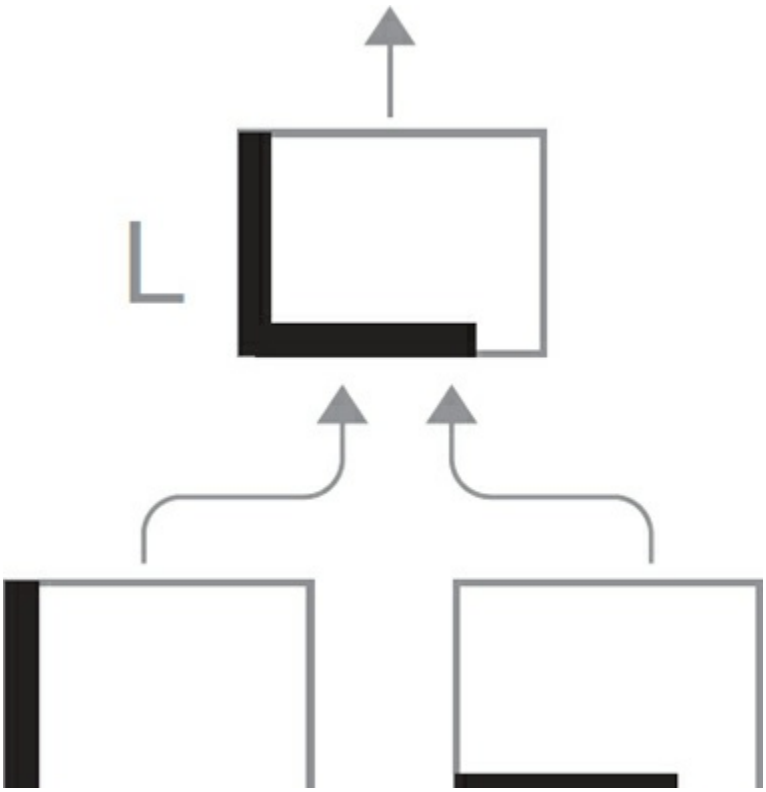
**Dos patrones diferentes, ambos elementos de “A”, y dos patrones diferentes de un nivel superior (“APPLE” y “PEAR”) del cual la “A” forma parte.**

“P”:



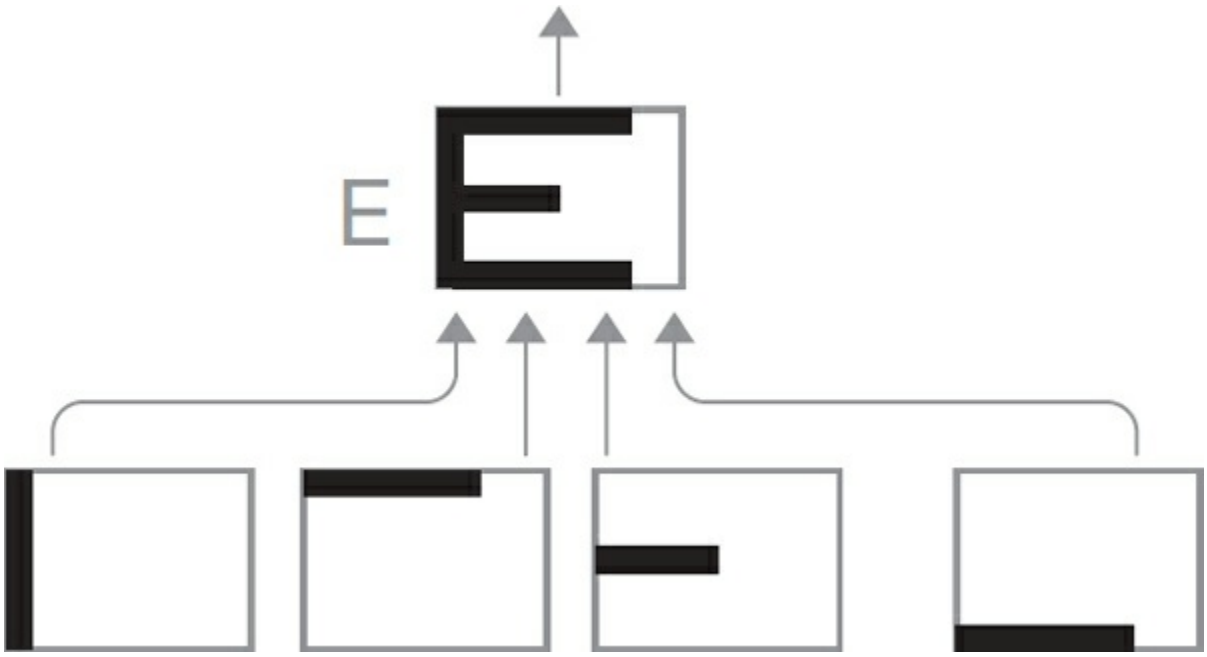

**Patrones que forman parte del patrón de nivel más alto “P”.**

“L”:



**Patrones que forman parte del patrón de nivel más alto "L".**

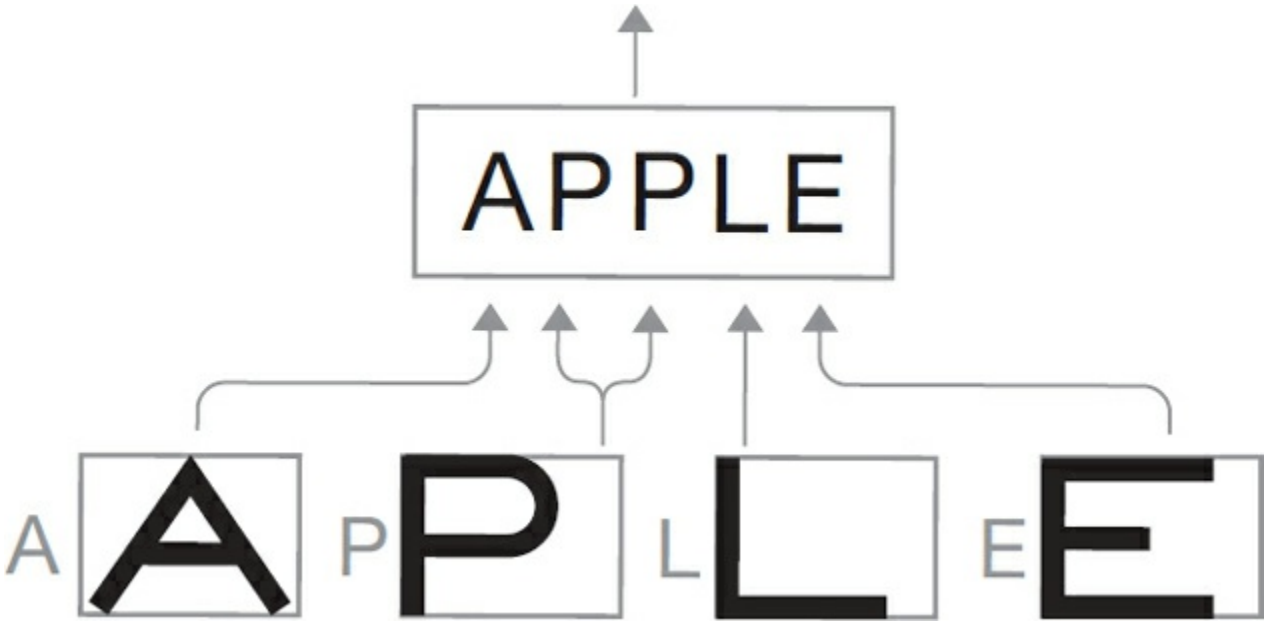
"E":




**Patrones que forman parte del patrón de nivel más alto "E".**

Estos patrones de letras alimentan un patrón de categoría todavía más alta llamado palabras. (La palabra "palabras" es nuestra categoría lingüística para este concepto, sin embargo el neocórtex las trata solo como patrones).

"APPLE":




En un lugar diferente del córtex, en lugar de letras impresas, se procesan *imágenes* reales de objetos mediante una jerarquía de patrones similar. Si usted mira una manzana real, los reconocedores de nivel bajo detectarán los patrones de los

contornos curvos y del color de la superficie, cosa que provoca que un reconocedor de patrones dispare su axón como diciendo: “oye, chicos, acabo de ver una manzana de verdad”. Asimismo, otros reconocedores de patrones detectarán combinaciones de frecuencias de sonido que hacen que sea posible que un reconocedor de patrones en el córtex auditivo dispare su axón indicando: “acabo de oir la palabra hablada ‘manzana’”.

Por otra parte, no hay que olvidar el factor de la redundancia. Para “manzana” en cada una de sus formas (escrita, hablada, vista) no poseemos un único reconocedor de patrones. Es posible que haya cientos de dichos reconocedores disparándose, si no más. La redundancia no solo aumenta la posibilidad de reconocer satisfactoriamente cada instancia de una manzana, sino que también está relacionada con las variaciones de las manzanas del mundo real. Así, para los objetos que reciban el nombre de manzana existen reconocedores de patrones que se corresponden con las muchas formas que toman las manzanas, según sus perspectivas, colores, sombras, formas y variedades diferentes.

También tenga en cuenta que la jerarquía mostrada anteriormente es una jerarquía d e *conceptos*. Debido a lo fina que es la estructura del neocórtex, estos reconocedores no se encuentran situados físicamente unos encima de otros, su altura física es la de un único reconocedor de patrones. Por tanto, la jerarquía conceptual surje mediante las conexiones entre los reconocedores de patrones individuales.

Un atributo importante de la PRTM* es la forma en la que los reconocimientos tienen lugar en el interior de cada módulo de reconocimiento de patrones. Almacenado en cada módulo existe un peso correspondiente a cada dendrita de entrada que indica cuál es la importancia de dicha entrada a la hora de llevar a cabo el reconocimiento. El reconocedor de patrones posee un umbral de disparo que indica que dicho reconocedor de patrones ha reconocido satisfactoriamente el patrón del que es responsable. No obstante, no todos y cada uno de los patrones de entrada tienen que estar presentes para que un reconocedor se dispare. Es probable que el reconocedor se disparare si falta una entrada de poco peso, pero es menos probable que lo haga si la que falta es una entrada de alta importancia. Cuando se dispara, un reconocedor de patrones lo que viene a decir es: “el patrón del que soy responsable es probable que esté presente”.

El reconocimiento satisfactorio de un patrón por parte de un módulo va más allá del mero recuento de las señales de entrada que están activadas (incluso si este recuento fuera medido por parámetros de importancia). El tamaño de cada entrada cuenta. Así, para cada entrada hay otro parámetro que indica el tamaño esperado de la entrada, e incluso otro que indica lo variable que es dicho tamaño. Para entender cómo funciona esto, supongamos que tenemos un reconocedor de patrones que es el responsable del reconocimiento de la palabra hablada “steep”*. Esta palabra hablada consta de cuatro sonidos: [s], [t], [E] y [p]. El fonema [t] es lo que se

conoce como “consonante dental”, que significa que se crea mediante el estallido de ruido que hace la lengua cuando el aire cesa de estar en contacto con los dientes superiores. En principio, la pronunciación lenta del fonema [t] es imposible. El fonema [p] está considerado como una “consonante oclusiva” u “oclusiva oral”, lo que significa que se crea cuando el tracto vocal es bloqueado repentinamente (en el caso de la [p] por parte de los labios), de manera que el aire deja de pasar. Necesariamente, tiene que ser rápida. La vocal [E] viene producida por las resonancias de las cuerdas vocales y de la boca abierta. Está considerada como una “vocal larga”, lo que significa que perdura durante un periodo de tiempo mucho más largo que otras consonantes tales como [t] y [p]. Sin embargo, su duración puede variar bastante. El fonema [s] es conocido como una “consonante sibilante” y viene producida por el choque del aire contra el filo de los dientes, que se mantienen cerrados. Por lo general, su duración es más corta que la de una vocal larga como la [E], pero esto también varía (en otras palabras, la [s] puede ser dicha rápidamente o alargarse).

Gracias a nuestro trabajo en el campo del reconocimiento del habla, hemos descubierto que para reconocer patrones del habla es necesario codificar este tipo de información. Por ejemplo, las palabras “step”** y “steep” son muy similares. Aunque el fonema [e] de “step” y el fonema [E] de “steep” son en cierto modo diferentes sonidos vocálicos (ya que poseen frecuencias reverberantes diferentes), no es fiable distinguir estas dos palabras basándose en estos sonidos vocales que tan a menudo dan lugar a confusión. Es mucho más fiable considerar el hecho de que la [e] en “step” es, comparada con la [E] de “steep”, relativamente breve.

Podemos codificar este tipo de información con dos números por cada entrada: un número para el tamaño esperado y un número para el grado de variabilidad de dicho tamaño. En la palabra “steep” de nuestro ejemplo, tanto [t] como [p] tienen una expectativa de duración muy corta, así como una corta expectativa de variabilidad (es decir, no esperamos escuchar largas ‘t’ y ‘p’). El sonido [s] tendría una expectativa de duración corta, pero una variabilidad más larga, ya que es posible arrastrarla. Por su parte, el sonido [E] posee una duración esperada larga, así como un alto grado de variabilidad.

En los ejemplos hablados, el parámetro “tamaño” se refiere a la duración. Sin embargo, el tiempo solo representa una de las posibles dimensiones. En el transcurso de nuestro trabajo sobre el reconocimiento de caracteres descubrimos que la información espacial era igualmente importante para reconocer las letras impresas (por ejemplo, el punto sobre la letra “i” se espera que sea mucho más pequeño que la parte situada debajo del punto). A niveles mucho más elevados de abstracción, el neocórtex maneja patrones a todo tipo de escalas*, tales como niveles de atractivo, ironía, felicidad, frustración y muchísimos más. Así, podemos encontrar similitudes

entre escalas** bastante diferentes, tal y como hizo Darwin cuando relacionó el tamaño físico de los cañones geológicos con el grado de diferenciación entre especies.

En un cerebro biológico, el origen de estos parámetros se encuentra en las experiencias del propio cerebro. No nacemos con un conocimiento innato de los fonemas, de hecho, los diferentes lenguajes poseen diferentes conjuntos de fonemas. Esto implica que los diferentes ejemplos de un patrón son codificados mediante los parámetros que cada reconocedor de patrones ha aprendido (ya que se necesitan múltiples instancias de un patrón para determinar la esperada distribución de magnitudes correspondiente a los inputs que le llegan al patrón). En algunos sistemas de IA, estos tipos de parámetros son codificados manualmente por expertos (por ejemplo, lingüistas que nos pueden decir las duraciones esperadas de los diferentes fonemas, tal y como he explicado anteriormente). Durante mi propio trabajo, descubrimos que hacer que un sistema de IA descubriese estos parámetros por sí mismo a partir de datos de entrenamiento (de forma similar a cómo lo hace el cerebro) significaba dar un enfoque superior al problema. A veces usábamos un enfoque híbrido, es decir, dotábamos al sistema con la intuición de los expertos humanos en lo que concernía a los ajustes iniciales de los parámetros y luego hacíamos que el sistema de IA refinase automáticamente estas estimaciones mediante un proceso de aprendizaje basado en ejemplos reales del habla.

Lo que hace el módulo de reconocimiento de patrones es calcular las probabilidades (es decir, la probabilidad según todas las experiencias previas) de que el patrón responsable del reconocimiento se vea realmente representado por sus inputs activos. Cada input que llega al módulo está activo si se dispara el correspondiente reconocedor de patrones de nivel más bajo (cosa que significa que el patrón de nivel más bajo ha sido reconocido). Cada input también codifica el tamaño observado en algunas de las dimensiones pertinentes, como por ejemplo la duración temporal, la magnitud física o alguna otra escala***, de manera que el tamaño pueda ser comparado por el módulo (incluyendo los parámetros del tamaño almacenados para cada input) mediante el cálculo de la probabilidad general del patrón.

¿Cómo calcula el cerebro (y cómo puede hacer lo mismo un sistema de IA) la probabilidad general de que un patrón (el módulo responsable del reconocimiento) esté presente dados (1) los inputs (a cada uno de los cuales le corresponde un tamaño), (2) los parámetros almacenados que tienen que ver con el tamaño (el tamaño esperado y la variabilidad del tamaño) correspondientes a cada input y (3) los parámetros sobre la importancia de cada input? En las décadas de 1980 y 1990, otros y yo fuimos pioneros a la hora de desarrollar un método matemático llamado modelos jerárquicos ocultos de Márkov para estudiar estos patrones y luego usarlos para reconocer patrones jerárquicos. Utilizamos esta técnica en el reconocimiento

del habla humana, así como para la comprensión del lenguaje natural. En el capítulo 7 describo este enfoque con mayor profundidad.

Volviendo al flujo de reconocimiento desde un nivel de reconocedores de patrones hasta el siguiente, en el ejemplo anterior vemos cómo la información fluye hacia arriba en la jerarquía conceptual desde las características básicas de las letras, pasando por las letras y llegando hasta las palabras. Desde ahí, los reconocimientos continúan fluyendo hacia arriba hasta llegar a las frases y luego hasta estructuras del lenguaje más complejas. Si subimos varias docenas de niveles más, llegamos hasta los conceptos de nivel más alto como la ironía o la envidia. Aunque todos los reconocedores de patrones trabajan simultáneamente, los reconocimientos necesitan tiempo para escalar por esta jerarquía conceptual. Así, el recorrido de cada nivel tarda entre unas pocas centésimas hasta unas pocas décimas de segundo en ser procesado. Hay experimentos que han demostrado que el patrón de un nivel moderadamente alto como el de una cara tarda por lo menos una décima de segundo. Si se dan distorsiones importantes, puede tardar un segundo entero. Si el cerebro fuera secuencial (como los son los ordenadores convencionales) y realizara cada reconocimiento de patrones secuencialmente, se vería obligado a sopesar cada uno de los patrones de bajo nivel posibles antes de continuar hasta el siguiente nivel. Así, para atravesar cada patrón se necesitarían millones de ciclos. Exactamente esto es lo que ocurre cuando simulamos estos procesos en un ordenador. Sin embargo, hay que tener en cuenta que los ordenadores procesan millones de veces más rápido que nuestros circuitos biológicos.

Es muy importante señalar que, al igual que la información fluye hacia arriba por la jerarquía conceptual, también lo hace hacia abajo. Este flujo hacia abajo es más importante si cabe. Si, por ejemplo, estamos leyendo de izquierda a derecha y ya hemos visto y reconocido las letras “M”, “A”, “N”, “Z”, *„A”* y “N”, el reconocedor de “MANZANA” predecirá que es probable encontrar una “A” en la siguiente posición. El reconocedor enviará una señal *hacia abajo* hasta el reconocedor de la “A” y vendrá a decir: “por favor, estate atento, hay una alta probabilidad de que veas tu patrón ‘A’ muy pronto, así que búscalo”. Entonces, el reconocedor de la “A” ajusta su umbral de manera que el reconocimiento de una “A” sea más probable. Así, si a continuación aparece una imagen que se parezca vagamente a una “A”, pero que sea lo suficientemente borrosa como para que no hubiera sido reconocida como una “A” en circunstancias “normales”, el reconocedor “A”, ya que la estaba esperando, es posible que sin embargo indique que efectivamente ha visto una “A”.

Por tanto, el neocórtex predice lo que espera encontrarse. Prever el futuro es una de las razones primordiales por las que poseemos un neocórtex. En el nivel conceptual más alto estamos continuamente haciendo predicciones (quién va a ser el próximo en atravesar la puerta, qué es lo siguiente que seguramente va a decir

alguien, qué es lo que esperamos ver cuando doblemos la esquina, los posibles resultados de nuestras propias acciones, etc.). Estas predicciones tienen lugar constantemente en *todos y cada uno* de los niveles de la jerarquía del neocórtex. No obstante, a menudo hay personas, cosas y palabras a las que reconocemos erróneamente debido a que nuestro umbral para confirmar un patrón esperado es demasiado bajo.

Además de señales positivas, también se dan señales negativan o inhibidoras que indican que un patrón determinado es menos probable que exista. Estas pueden provenir de niveles conceptuales inferiores (por ejemplo, el reconocimiento de un bigote inhibirá la posibilidad de que una persona que estoy viendo en la cola de salida sea mi mujer), o de un nivel superior (por ejemplo, sé que mi mujer está de viaje, de manera que la persona de la cola de salida no puede ser ella). Cuando un reconocedor de patrones recibe una señal inhibidora, el umbral de reconocimiento disminuye, pero sigue siendo posible que el patrón se dispare (de manera que si la persona de la cola es realmente ella, es posible que de todas maneras la reconozca).

## *La naturaleza del flujo de datos que llega hasta los reconocedores de patrones neocorticales*

Consideremos con más más profundidad el aspecto de los datos correspondientes a un patrón. Si el patrón es una cara, los datos existen en al menos dos dimensiones. No podemos decir que los ojos sean necesariamente lo que venga antes, que los siga la nariz, etc. Lo mismo es cierto para la mayoría de los sonidos. Una obra musical posee por lo menos dos dimensiones. Es posible que haya más de un instrumento y/o voz que esté emitiendo sonidos al mismo tiempo. Y no solo eso, una nota perteneciente a un instrumento complejo como por ejemplo el piano se compone de múltiples frecuencias. Una voz humana se compone de varios niveles simultaneos de energía en docenas de bandas de frecuencia diferentes, de manera que un patrón de sonido puede volverse complejo en cualquier momento y además estos momentos complejos se extienden en el tiempo. Los inputs táctiles también son bidimensionales, ya que la piel es un órgano sensitivo bidimensional. Además, dichos patrones pueden cambiar durante la tercera dimensión que viene representada por el tiempo.

Así, daría la impresión de que el input que llega a un pocresador de patrones del neocórtex tuviera que constar de patrones de dos (si no tres) dimensiones. Sin embargo, en la estructura del neocórtex podemos observar que los inputs de los patrones son solo listas unidimensionales. Todo nuestro trabajo en el campo de la creación de sistemas artificiales de reconocimiento de patrones (tales como sistemas

de reconocimiento visual y reconocimiento del habla) demuestra que podemos (y de hecho así lo hicimos) representar fenómenos bi y tridimensionales mediante una de dichas listas unidemsionales. Describiré cómo funcionan estos métodos en el capítulo 7, pero por ahora podemos quedarnos con el hecho de que el input llega a todos y cada uno de los procesadores de patrones en una lista unidimensional, aunque el patrón mismo pueda reflejar intrínsicamente más de una dimensión.

En este punto deberíamos mencionar el hecho de que los patrones que hemos aprendido a reconocer (por ejemplo, un perro en concreto o la idea general de un "perro", una nota o una composición musical) representan exactamente el mismo mecanismo en el que se fundamentan nuestros recuerdos. De hecho, nuestros recuerdos son patrones organizados a modo de listas donde cada ítem de cada lista es otro patrón en la jerarquía cortical que hemos aprendido y posteriormente reconocido cuando se nos ha hecho presente mediante el estímulo apropiado. De hecho, las memorias existen en el neocórtex para ser reconocidas.

La única excepción a todo esto se da en el nivel conceptual más bajo posible, en el cual los datos del input correspondientes a un patrón representan información sensorial específica (por ejemplo, datos por imágenes procedentes del nervio óptico). Sin embargo, incluso este patrón de nivel más bajo, en el momento de alcanzar el córtex, ya ha sufrido una transformación importante en patrones simples. Las listas de patrones que constituyen un recuerdo se ordenan de atrás a adelante, por lo que solo podemos rememorar nuestros recuerdos en dicho orden y de ahí nuestras dificultades a la hora de invertir el orden de nuestros recuerdos.

Un recuerdo tiene que ser desencadenado por otro pensamiento/recuerdo (son lo mismo). Al percibir un patrón podemos experimentar este mecanismo detonante. Al percibir "M", "A", "N", "Z", "A" y "N", el patrón de "M A N Z A N A" predice que veremos una "A" y desencadena el patrón "A", que es el que se espera en este momento. Por tanto, nuestro córtex está "pensando" en ver una "A" incluso antes de que la veamos. Si esta interacción concreta de nuestro córtex recibe nuestra atención, pensaremos sobre la "A" antes de verla o incluso aunque no llegemos a verla nunca. Un mecanismo similar es el que desencadenan los viejos recuerdos. Normalmente se produce toda una cadena de vínculos similares a estos. Incluso teniendo algún nivel de conciencia sobre los recuerdos (es decir, los patrones) que desencadenaron el viejo recuerdo, los recuerdos (patrones) no están dotados ni de lenguaje ni de etiquetas. Esta es la razón por la que puede parecer que viejos recuerdos se nos hacen presentes de repente. Después de haberlos enterrado y no haberlos activado quizás desde hace años, necesitan un desencadenante, de la misma manera que una página web necesita de un enlace web para ser activada. E igual que una página web puede volverse "huérfana" debido a que ninguna otra página enlaza con ella, lo mismo puede pasarle a nuestros recuerdos.

En su mayor parte, nuestros pensamientos se activan de uno de estos dos modos:

indirectamente o directamente; y ambos modos utilizan los mismos enlaces corticales. En el modo indirecto, permitimos que los enlaces se desarrollen sin intentar dirigirlos hacia una dirección en particular. Algunas formas de meditación (por ejemplo la meditación transcendental, de la cual yo soy practicante) se basan en permitir que la mente haga exactamente esto. Asimismo, los sueños también poseen esta cualidad.

En el pensamiento dirigido intentamos pasar por un proceso más ordenado en el que se rememora un recuerdo (por ejemplo, una historia) o se resuelve un problema. Esto también supone transistar por listas en nuestro neocórtex. Aun así, el fresnesí menos estructurado del pensamiento no dirigido también acompañará a este proceso. Por tanto, el contenido completo de nuestro pensamiento está muy desordenado, un fenómeno que James Joyce ilustró en sus novelas mediante el llamado "monólogo interior" o "flujo de conciencia"*.

Cuando se repasan los recuerdos/ historias/ patrones de la vida, ya tengan que ver estos con el encuentro fortuito con una madre que empuja un carrito de bebé porque está dando una vuelta o con una historia más importante como por ejemplo la manera en la que usted conoció a su esposa, los recuerdos se componen de una secuencia de patrones. Dado que estos patrones no vienen etiquetados ni con palabras, ni con sonidos, ni con vídeos, al tratar de rememorar un hecho significativo lo que esencialmente se realiza es una reconstrucción mental de imágenes, ya que las imágenes reales no existen.

Si pudiéramos "leer" la mente de alguien y asomarnos a lo que está pasando exactamente en el neocórtex, sería muy difícil interpretar los recuerdos de dicha persona. Independientemente de que lo que miráramos fueran los patrones que simplemente están almacenados en el neocórtex a la espera de ser disparados o aquellos que han sido disparados y actualmente están siendo vividos como pensamientos activos, lo que "veríamos" es la activación simultánea de millones de reconocedores de patrones. Una centésima de segundo después veríamos un conjunto diferente compuesto por un número similar de reconocedores de patrones activados. Cada uno de estos patrones sería una lista de otros patrones, y estos serían a su vez una lista de otros patrones, y así sucesivamente hasta que llegáramos a los patrones simples más elementales posibles en el nivel más bajo. Sería extraordinariamente difícil interpretar lo que estos patrones de nivel superior significan si no copiáramos *toda* la información contenida en cada nivel de nuestro propio córtex. Así, cada patrón en nuestro neocórtex solo tiene significado si se tiene en cuenta toda la información contenida en los niveles que están por debajo de él. Además, otros patrones pertenecientes al mismo nivel y superiores también son importantes a la hora de interpretar un patrón en particular, ya que proporcionan información sobre el contexto. Por tanto, una verdadera lectura de la mente no solo requeriría la detección

de las activaciones de los axones importantes en el cerebro de una persona, sino que, para entender estas activaciones, también se necesitaría examinar el neocórtex por completo, incluyendo todos sus recuerdos.

Cuando experimentamos nuestros propios pensamientos y recuerdos "sabemos" lo que significan, pero no se presentan como pensamientos y rememoraciones inmediatamente explicables. Si los queremos compartir con los demás, tenemos que traducirlos en lenguaje. Esta tarea también la realiza el neocórtex. Para ello utiliza reconocedores de patrones entrenados mediante patrones que hemos aprendido con el objetivo de usar el lenguaje. El propio lenguaje es en sí muy jerárquico y ha evolucionado para aprovecharse de la naturaleza jerárquica del neocórtex, que a su vez refleja la naturaleza jerárquica de la realidad. La innata capacidad de los humanos para aprender las estructuras jerárquicas del lenguaje a las que Noam Chomsky hace referencia, refleja la estructura del neocórtex. En 2002, en un estudio del que es co-autor, Chomsky se refiere al atributo de la "recursión" como responsable de la capacidad para el lenguaje que es única de los humanos (4). Según Chomsky, la recursión es la capacidad para juntar pequeñas partes y formar una porción más grande para luego utilizar dicha porción como parte componente de otra estructura y repetir este proceso de forma iterativa. Así es como, partiendo de un conjunto limitado de palabras, somos capaces de construir las elaboradas estructuras de las frases y de los párrafos. Aunque Chomsky no se estaba refieriendo explícitamente a la estructura del cerebro, la capacidad que describe es exactamente la función que realiza el neocórtex.

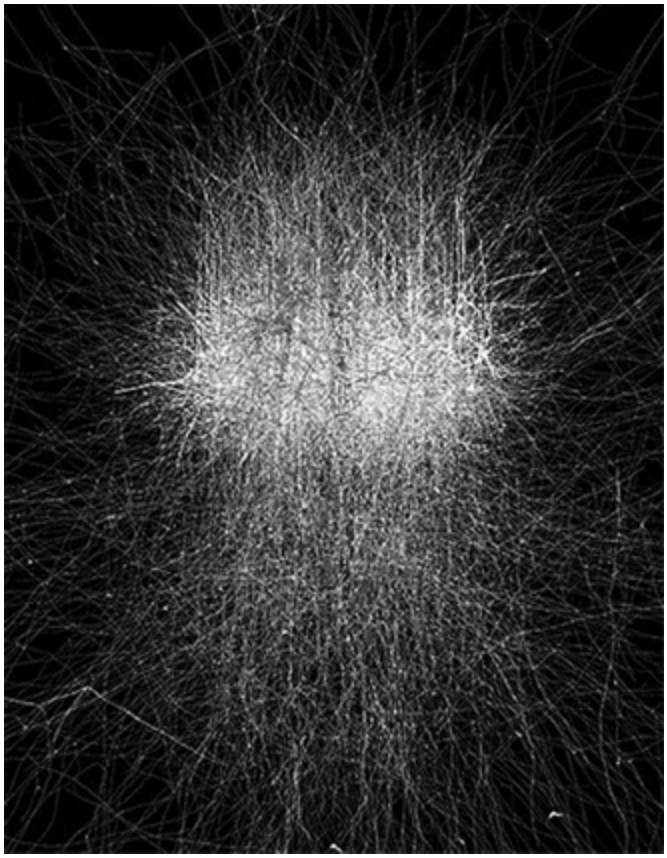
Especies inferiores de mamímeros suelen utilizar toda la capacidad de su neocórtex para afrontar los retos que les presenta su hábitat particular. La especie humana adquirió capacidades adicionales al desarrollar un córtex substancialemente mayor que le permitió manejar el lenguaje hablado y escrito. Algunas personas han aprendido estas capacidades mejor que otras. Si una historia en particular la hemos contado muchas veces, empezaremos a aprender la secuencia del lenguaje que describe la historia como una serie de secuencias separadas. Incluso en este caso, nuestra memoria no es una precisa secuencia de palabras, sino más bien de estructuras del lenguaje que tenemos que traducir en secuecias específicas de palabras cada vez que exponemos la historia. Esta es la razón por la cual contamos una historia de forma un poco diferente cada vez que la compartimos con los demás (a no ser que hayamos aprendido la secuencia de palabras exacta a modo de patrón).

Para cada una de estas descripciones correspondientes a procesos específicos de pensamiento también tenemos que tener en cuenta la cuestión de la redundancia. Como ya he mencionado anteriormente, no poseemos un único patrón que represente las entidades importantes de nuestras vidas, ya constituyan dichas entidades categorías sensoriales, conceptos lingüísticos o recuerdos sobre sucesos. Cada patrón importante (en todos los niveles) se repite muchas veces. Algunas

recurrencias representan meras repeticiones, mientras que muchas de ellas representan diferentes perspectivas y puntos de observación. Esta es la principal razón por la cual podemos reconocer la forma de una cara familiar desde varias direcciones y en condiciones de luminosidad diferentes. Además, cada nivel de la jerarquía posee una redundancia substancial, lo cual permite la variabilidad suficiente acorde con el concepto en cuestión.

Así, si nos imagináramos examinando su neocórtex cuando está mirando a una persona amada en particular, lo que veríamos sería una gran cantidad de disparos de axones pertenecientes a los reconocedores de patrones de cada nivel, desde el nivel básico de los patrones sensoriales primitivos hasta los muy distintos patrones que representan la imagen del ser querido. También veríamos enormes cantidades de disparos que representan otros aspectos de la situación, tales como los movimientos de la persona, lo que dice, etc. De modo que si esta experiencia le parece mucho más rica que un mero viaje de ascenso por una jerarquía de características, es que lo es.

Sin embargo, el mecanismo básico de ascenso por una jerarquía de reconocedores de patrones en la cual cada nivel conceptual superior representa un concepto más abstracto y más integrado sigue siendo válido. El flujo de información hacia abajo es todavía mayor, ya que cada nivel activado de un patrón reconocido envía predicciones al siguiente reconocedor de patrones de nivel más bajo sobre lo que es probable que se vaya a encontrar a continuación. La aparente exuberancia de las experiencias humanas es el resultado de que todos los cientos de millones de reconocedores de patrones de nuestro neocórtex estén considerando sus inputs simultáneamente.



**Una simulación por ordenador del disparo simultáneo de muchos reconocedores de patrones en el neocórtex.**

En el capítulo 5 trataré el flujo de información perteneciente al tacto, a la vista, al oído y a otros órganos sensoriales en el interior del neocórtex. Estos primeros inputs son procesados por regiones corticales dedicadas a los tipos relevantes de inputs sensoriales (aunque la tarea de estas regiones está sujeta a una enorme plasticidad que refleja la uniformidad básica en cuanto a la función del neocórtex). La jerarquía conceptual continúa más allá de los conceptos más elevados de cada región sensorial del neocórtex. Así, las áreas de asociación integran input procedente de diferentes inputs sensoriales. Cuando oímos algo que quizá suena como la voz de nuestra esposa y luego vemos algo que quizá indique su presencia, no nos enfrascamos en un elaborado proceso de deducción lógica. En vez de eso, percibimos instantáneamente que nuestra esposa está presente partiendo de la combinación de estos reconocimientos sensoriales. Integramos todas las indicaciones sensoriales y perceptuales relevantes (quizás incluso el olor de su perfume o de su colonia) a modo de percepción de múltiples niveles.

A un nivel conceptual situado por encima de las áreas de asociación sensoriales del córtex somos capaces de ocuparnos (percibiéndolos, rememorándolos y reflexionándolos) de conceptos todavía más abstractos. En el nivel más alto reconocemos patrones tales como *eso tiene gracia, ella es guapa, eso es irónico*, etc. Nuestros recuerdos también incluyen estos abstractos patrones de reconocimiento. Por ejemplo, podemos recordar que estuvimos dando un paseo con

alguien y que ella dijo algo gracioso de lo que nos reímos, aunque es posible que no nos acordemos del propio chiste. Simplemente, la secuencia del recuerdo perteneciente a esa remembranza grabó la percepción del humor pero no el contenido preciso de lo que resultó gracioso.

En el capítulo anterior indiqué que a menudo podemos reconocer un patrón aunque no lo reconozcamos lo suficientemente bien como para ser capaces de describirlo. Por ejemplo, yo creo que podría identificar, de entre un grupo de fotos de otras mujeres, una foto de la mujer con el carrito de bebé que ví hoy, pese al hecho de que no soy capaz de visualizarla y de que no pueda describirla muy específicamente. En este caso, mi recuerdo de ella es una lista de ciertas características de alto nivel. Estas características no poseen etiquetas lingüísticas o visuales asociadas a ellas y no se trata de imágenes pixeladas, de manera que, aunque soy capaz de pensar en ella, no soy capaz de describirla. Sin embargo, si se me pone delante una foto de ella puedo procesar la imagen, lo cual desemboca en el reconocimiento de las mismas características de alto nivel que fueron reconocidas la primera vez que la vi. De esta manera sería capaz de determinar que las características encajan y por tanto elegir su foto confiando en no equivocarme.

Aunque durante el paseo solo vi a esta mujer una vez, es probable que ya existan en mi neocórtex múltiples copias de su patrón. Sin embargo, si no pienso en ella durante un cierto tiempo estos reconocedores de patrones serán reasignados a otros patrones. Por esta razón los recuerdos, con el tiempo, se vuelven más tenues, ya que la cantidad de redundancia se reduce hasta que ciertos recuerdos se extinguen. Sin embargo, ahora que ya he rememorado a esta mujer en particular escribiendo aquí sobre ella, es probale que no la vaya a olvidar fácilmente.

## *Autoasociación e invarianza*

En el capítulo anterior expuse cómo podemos reconocer un patrón incluso si el patrón completo no está presente o si está distorsionado. La primera de estas capacidades se llama autoasociación: la capacidad de asociar un patrón con una parte de sí mismo. Intrísicamente, la estructura de todos los reconocedores de patrones sustenta esta capacidad.

Cuando los inputs procedentes de un reconocedor de patrones de nivel más bajo fluyen hacia un reconocedor de nivel más alto, la conexión que se establece puede poseer un “peso” que indique la importancia que dicho elemento en particular tiene en el patrón. Así, los elementos más siginificativos de un patrón tienen un mayor peso a la hora de tomar en consideración si dicho patrón debería dispararse bajo la categoría de “reconocido”. Es altamente probable que la barba de Lincoln, las

patillas de Elvis y el famoso gesto con la lengua de Einstein tengan un gran peso en los patrones que hemos aprendido sobre la apariencia de estos iconos. El reconocedor de patrones calcula una probabilidad que toma en consideración los parámetros de importancia. De este modo, la probabilidad general es más pequeña si uno o varios de los elementos faltan, aunque el umbral de reconocimiento puede que siga siendo superado. Tal y como señalé, el cálculo de la probabilidad general de que el patrón esté presente es más complicado que la simple suma de los pesos, ya que los parámetros del tamaño también tienen que ser tenidos en cuenta. Si el reconocedor de patrones ha recibido una señal procedente de un reconocedor de nivel más alto, dicho patrón es "esperado", por lo que el umbral se hace más grande (es decir, es más fácil rebasarlo). No obstante, una señal así puede que simplemente sea añadida al peso total de los inputs, compensando con ello la falta del elemento. Esto pasa en todos los niveles, de manera que un patrón como por ejemplo una cara que se encuentra varios niveles por encima del nivel más bajo puede ser reconocido aunque falten varias de sus características.

La capacidad para reconocer patrones incluso cuando ciertos aspectos de los mismos han sido transformados recibe el nombre de invarianza, y es abordada de cuatro maneras diferentes.

Primero, existen transformaciones globales que son realizadas antes de que el neocórtex reciba datos sensoriales. En la página 90, en la sección "La vía sensitiva", expondremos el viaje realizado por los datos sensoriales a partir de los ojos, los oídos y la piel.

El segundo método se aprovecha de la redundancia de nuestra memoria cortical de patrones. Sobre todo en lo que se refiere a ítems importantes, hemos aprendido muchas perspectivas y puntos de vista diferentes para cada patrón. Así, muchas variaciones son almacenadas y procesadas por separado.

El tercer y más poderoso método consiste en la capacidad para combinar dos listas. Una lista puede tener un conjunto de transformaciones que hayamos aprendido que puede que sean aplicables a una cierta categoría de patrón. El córtex aplicará esta misma lista de posibles cambios en otro patrón. Así es cómo comprendemos tropos tales como las metáforas y los símiles.

Por ejemplo, hemos aprendido que ciertos fonemas (los sonidos básicos del lenguaje) pueden estar ausentes en el lenguaje hablado (por ejemplo en la palabra "parao"*). Si posteriormente aprendemos una nueva palabra hablada (por elemplo, "demasiao"**), seremos capaces de reconocer dicha palabra aunque falte uno de sus fonemas incluso si no hemos escuchado anterioremente esa palabra bajo dicha forma, ya que nos hemos familiarizado con el fenómeno general de que ciertos fonemas sean omitidos. Pongamos otro ejemplo. Podemos aprender que a un artista en particular le gusta enfatizar, mediante el aumento de su tamaño, ciertos elementos de una cara, como por ejemplo la nariz. De esta manera, podemos identificar una cara con la que

estamos familiarizados a la que se le ha aplicado dicha modificación, incluso sin haber visto antes dicha modificación en dicha cara. Así, ciertas modificaciones artísticas enfatizan exactamente las mismas características que son reconocidas por el reconocimiento de patrones en el que se basa nuestro neocórtex. Como ya he dicho, precisamente en esto se basan las caricaturas.

El cuarto método se deriva del tamaño de los parámetros que permiten que un módulo individual codifique múltiples instancias de un patrón. Por ejemplo, hemos oído la palabra "steep"*** muchas veces. Un determinado módulo de reconocimiento de patrones que reconozca esta palabra hablada puede codificar estos múltiples ejemplos indicando que la duración de la [E] posee una variabilidad altamente esperada. Si todos los módulos de las palabras que incluyen la [E] compartieran un fenómeno similar, dicha variabilidad podría ser codificada en los modelos de la propia [E]. Sin embargo, las diferentes palabras que incorporan la [E] (o muchos otros fonemas) pueden poseer diferentes cantidades de variabilidad esperada. Por ejemplo, la palabra "peak"* es probable que no tenga un fonema [E] tan prolongado como lo tiene la palabra "steep".

## *Aprendizaje*

> ¿Acaso no estamos creando a nuestros sucesores en la supremacía sobre la Tierra? ¿No lo hacemos añadiendo día a día más bellaza y delicadeza a su organización, dotándoles día a día de más capacidades y suministrandoles más y más del poder de automatismo y de la autoregulación que acabará siendo mejor que cualquier intelecto?
>
> —SAMUEL BUTLER, 1871
>
> Las actividades más importantes llevadas a cabo por los cerebros realizan cambios sobre sí mismas".
>
> —MARVIN MINSKY, *THE SOCIETY OF MIND*

Hasta ahora hemos examinado cómo reconocemos patrones sensoriales y perceptivos, y cómo rememoramos secuencias de patrones (nuestros recuerdos sobre cosas, gente y sucesos). Sin embargo, no nacemos con un neocórtex lleno de estos patrones. Cuando nuestro cerebro se crea, es territorio virgen. Tiene la capacidad de aprender y por tanto de crear conexiones entre sus reconocedores de patrones, pero dichas conexiones las consigue a partir de la experiencia.

Este proceso de aprendizaje comienza incluso antes de que hayamos nacido y tiene lugar al mismo tiempo que el proceso biológico de desarrollo del cerebro. Un feto de un mes ya tiene cerebro, aunque en su mayor parte es un cerebro reptiliano, ya que en el útero el feto recrea a alta velocidad la evolución biológica. En el momento del nacimiento, el cerebro es claramente humano y está dotado de un neocórtex humano a partir del tercer trimestre de gestación. En ese momento el feto

ya tiene experiencias y el neocórtex aprende. El bebé puede oir sonidos, sobre todo el latido del corazón de la madre, lo que seguramente explique por qué las cualidades rítmicas de la música son universales en toda cultura humana. Todas las civilizaciones humanas descubiertas han incorporado la música como parte de su cultura, cosa que no pasa con otras formas artísticas, como por ejemplo la pintura. También es verdad que el ritmo de la música se puede comparar a nuestra frecuencia cardiaca. Ciertamente, los ritmos musicales varían, si no fuera así la música no despertaría nuestro interés, pero las pulsaciones del corazón también varían (un latido demasiado regular es síntoma de enfermedad cardiaca). Los ojos de un feto se abren parcialmente 26 semanas después de la concepción y tras 20 semanas después de la concepción están completamente abiertos prácticamente todo el tiempo. Es posible que en el interior del útero no haya mucho que ver, sin embargo existen patrones de luz y oscuridad que el neocórtex comienza a procesar.

Así, aunque un bebé recién nacido ha tenido unas pocas experiencias en el interior del útero, estas son claramente limitadas. El neocórtex también puede aprender del cerebro antiguo (un tema que expondré en el capítulo 5), pero por lo general cuando un bebé nace tiene mucho por aprender (todo lo que va desde los sonidos y las formas primitivos primarios, hasta las metáforas y el sarcasmo).

El aprendizaje es crucial para la inteligencia humana. Si pudiéramos modelizar y simular el neocórtex humano a la perfección (tal y como intenta hacer el proyecto Blue Brain) y todas las demás regiones cerebrales cuyo funcionamiento es necesario para el cerebro, tales como el hipocampo y el tálamo, este cerebro no podría hacer demasiadas cosas, igual que un bebé recién nacido tampoco las puede hacer, aparte de ser muy mono, lo cual es sin duda una adaptación fundamental para la supervivencia.

Aprendizaje y reconocimiento tienen lugar simultáneamente. Inmediatamente empezamos a aprender y en cuanto hemos aprendido un patrón empezamos a reconocerlo de inmediato. Continuamente, el neocórtex intenta dotar de sentido al input que recibe. Si un nivel concreto no es capaz de procesar y de reconocer un patrón en su totalidad, dicho patrón es enviado hasta el siguiente nivel más alto. Si ninguno de los niveles tiene éxito a la hora de reconocer un patrón, éste es considerado como un patrón nuevo. El clasificar un patrón como nuevo no significa necesariamente que todos sus aspectos sean nuevos. Si observamos los cuadros de un artista en particular y vemos la cara de un gato con nariz de elefante, no seremos capaces de identificar cada una de sus diferentes características, pero nos daremos cuenta de que este patrón combinado se trata de algo novedoso y seguramente lo recordaremos. Los niveles conceptuales más altos del neocórtex, aquellos que comprenden el contexto (por ejemplo, el hecho de que este cuadro sea un ejemplo de las obras de un determinado artista y que estemos asistiendo a la inauguración de una exposición realizada por dicho artista), se darán cuenta de la inusual combinación de

patrones en la cara del gato-elefante, pero también incorporarán estos detalles contextuales a modo de patrones de memoria adicionales.

Nuevos recuerdos como por ejemplo la cara del gato-elefante son almacenados en un reconocedor de patrones válido para ello. El hipocampo juega un papel en este proceso y en el próximo capítulo expondremos lo que se conoce sobre este mecanismo biológico. En lo que respecta al modelo de nuesto neocórtex, basta con decir que los patrones que de otro modo no son reconocidos son almacenados como patrones nuevos, y son conectados adecuadamente a los patrones de nivel más bajo que los componen. Por ejemplo, la cara del gato-elefante se almacenará de diferentes maneras. La novedosa combinación de las partes faciales será almacenada, al igual que los recuerdos contextuales que incluyen al artista, a la situación y quizás al hecho de que nos riéramos la primera vez que la vimos.

Los recuerdos que sean reconocidos con éxito también pueden dar lugar a la creación de un nuevo patrón que consiga aumentar la redundancia. Por el contrario, si los patrones no son reconocidos perfectamente, es probable que sean almacenados como reflejo de una perspectiva diferente sobre el ítem reconocido.

Entonces, ¿cuál es el método general seguido para determinar qué patrones son almacenados? En términos matemáticos el problema puede ser expresado de la siguiente manera: haciendo uso de los límites disponibles para el almacenamiento de patrones, ¿cómo representar de forma óptima los patrones del input que hasta ahora hemos recibido? Aunque tiene sentido permitir una cierta cantidad de redundancia, no sería práctico llenar todo el área disponible para el almacenamiento (es decir, todo el neocórtex) con patrones repetidos, ya que eso no permitiría una diversidad suficiente de patrones. Un patrón como el fonema [E] de palabras habladas es algo que hemos experimentado incontables veces. Se trata de un patrón simple de frecuencias sonoras y sin duda alguna disfruta de una importante redundancia en nuestro neocórtex. Podríamos llenar todo nuestro neocórtex con patrones repetidos del fonema [E]. Sin embargo, la redundancia útil tiene un límite y un patrón corriente como este está claro que ya lo ha alcanzado.

Existe una solución matemática para este problema de optimización llamada programación lineal*, que busca soluciones teniendo en cuenta la mejor distribución posible de unos recursos limitados (en este caso, un número limitado de reconocedores de patrones) y que es representativa en todos aquellos casos en los que el sistema ha sido entrenado. La programación lineal está diseñada para sistemas con inputs unidimensionales. Esta es otra razón por la que se trata de una solución óptima para representar el input correspondiente a cada módulo de reconocimiento de patrones como una sucesión lineal de inputs. Se puede utilizar esta estrategia matemática en un sistema de software y, pese a que un cerebro real también se ve constreñido por las conexiones físicas de las que dispone para ser adaptadas entre

los reconocedores de patrones, el método no obstante es similar en ambos casos.

Una implicación importante de esta solución óptima es que las experiencias rutinarias son reconocidas, pero no dan como resultado la creación de un recuerdo permanente. En lo que respecta a mi paseo, experimenté millones de patrones a todos los niveles, desde contornos visuales básicos y sombras, hasta objetos tales como farolas, buzones, personas, animales y plantas de las que pasé al lado. Casi nada de lo que experimenté fue único y los patrones que reconocí hace mucho que han alcanzado su nivel óptimo de redundancia. El resultado es que no recuerdo casi nada de este paseo. Los pocos detalles de los que me acuerdo es probable que acaben siendo sustituidos los por nuevos patrones que se generarán después de dar otras pocas docenas de paseos y que serán escritos sobre los recuerdos de este paseo (a no ser por el hecho de que ya he memorizado este paseo en concreto al haber escrito sobre él).

Un punto importante que afecta tanto a nuestro neocórtex biológico como a nuestros intentos por emularlo es que resulta difícil aprender muchos niveles conceptuales simultáneamente. Básicamente, se pueden aprender uno o como mucho dos niveles conceptuales al mismo tiempo. Una vez que el aprendizaje es relativamente estable, se puede empezar a aprender el siguiente nivel. Así, se puede seguir ajustando (refinando) el aprendizaje en los niveles inferiores, pero el foco en el que se centra nuestro aprendizaje es el siguiente nivel de abstracción. Esto es cierto tanto al principio de la vida cuando somos recién nacidos que se enfrentan con contornos básicos, como más adelante cuando luchamos por aprender nuevas materias de estudio. Siempre pasamos de un nivel de complejidad al siguiente. El mismo fenómeno se da en las emulaciones tecnológicas del neocórtex. Sin embargo, si a estas se les presentan materiales cada vez más abstractos pertenecientes a niveles de complejidad consecutivos, las máquinas son capaces de aprender igual que los hacen los humanos, aunque todavía no con tantos niveles conceptuales como lo hacen los humanos.

El output de un patrón puede retroalimentar un patrón a un nivel más bajo o incluso al propio patrón, ya que el cerebro humano posee una capacidad recursiva muy poderosa. Un elemento de un patrón puede ser un punto de decisión basado en otro patrón. Esto es especialmente útil en el caso de las listas que conforman acciones, por ejemplo cuando se acude a otro tubo de pasta de dientes cuando el que se está usando está vacío. Estos condicionantes existen en todos los niveles. Tal y como saben todos aquellos que han intentado programar un procedimiento en un ordenador, los condicionantes son vitales para describir el curso de una acción.

# *El lenguaje del pensamiento*

El sueño actúa a modo de válvula de escape de un cerebro sobrecargado.

—SIGMUND FREUD. *LA INTERPRETACIÓN DE LOS SUEÑOS,* 1911

Cerebro: aparato con el que pensamos que pensamos.

—AMBROSE BIERCE, *THE DEVIL'S DICTIONARY*

Para resumir lo que hemos aprendido hasta ahora sobre el modo en el que el neocórtex funciona, acúdase por favor al diagrama del módulo de reconocimiento de patrones neocortical de la página 38.

a) Las dendritas acceden al módulo que representa al patrón. Aunque pueda parecer que los patrones tengan dos o tres cualidades dimensionales, vienen representados por una secuencia de señales unidimensional. El patrón tiene que hacerse presente según este orden secuencial para que el reconocedor de patrones sea capaz de reconocerlo. En último término, cada una de las dendritas está conectada a uno o más axones de los reconocedores de patrones de un nivel conceptual más bajo que han reconocido un patrón de nivel más bajo que forma parte de este patrón. Para cada uno de estos patrones de input pueden existir muchos reconocedores de patrones de nivel más bajo que pueden generar la señal que indica que el patrón de nivel más bajo ha sido reconocido. El umbral necesario para reconocer el patrón puede ser alcanzado aunque no todos los inputs hayan sido señalizados. El módulo calcula la probabilidad de que esté presente el patrón del cual es responsable. Este cálculo toma en consideración los parámetros de "importancia" y "tamaño" (véase la [f] de más abajo).

   Téngase en cuenta que algunas dendritas transmiten señales desde el interior del módulo y algunas lo hacen desde el exterior del mismo. Si todas las dendritas del input correspondientes a este reconocedor de patrones señalizan que, a excepción de uno o dos, sus patrones de nivel más bajo han sido reconocidos, entonces este reconocedor de patrones enviará una señal hacia abajo en dirección a los reconocedor(es) de patrones que reconocen los patrones de nivel más bajo que todavía no han sido reconocidos, y les indicará que hay una alta probablilidad de que pronto el patrón sea reconocido y que el/los reconocedor(es) de nivel más bajo deberían empezar a buscarlo.

b) Cuando, basándose en que todas o casi todas las señales de las dendritas del input se han activado, este reconocedor de patrones reconoce su patrón, el axón (output) de este reconocedor de patrones se activará. A su vez, este

axón puede conectarse con toda una red de dendritas que se conectan a muchos reconocedores de patrones de nivel más alto para los cuales este patrón realiza las veces de input. Esta señal transmite información sobre la magnitud, de manera que los reconocedores de patrones en el nivel conceptual inmediatamente superior puedan tomarlo en consideración.

c) Si un reconocedor de patrones de nivel más alto recibe una señal positiva procedente de todos o casi todos los patrones que lo conforman excepto de aquel representado por el reconocedor de patrones en sí, entonces ese reconocedor de patrones de nivel más alto es posible que envíe una señal hacia abajo hasta este reconocedor que indique que se está a la expectativa de su patrón. Dicha señal haría que este reconocedor de patrones aumentara su umbral, lo que significaría que sería más probable que mandase una señal a su axón indicando que se considera que su patrón ha sido reconocido aunque algunos de sus inputs falten o no estén claros.

d) Las señales inhibidoras procedentes de más abajo harían más difícil que este reconocedor de patrones reconozca su patrón. Esto puede ser el resultado del reconocimiento de patrones de nivel más bajo que no sean consistentes con el patrón asociado a este reconocedor de patrones (por ejemplo, el reconocimiento de un bigote por medio de un reconocedor de nivel más bajo haría menos probable que esta imagen fuese asociada con "es mi mujer").

e) Las señales inhibidoras procedentes de más arriba también hacen más difícil que este reconocedor de patrones reconozca su patrón. Esto puede ser el resultado de un contexto de nivel más alto que no sea congruente con el patrón asociado con este reconocedor.

f) A cada input le corresponden parámetros almacenados relativos a la importancia, el tamaño esperado y la variabilidad esperada en cuanto al tamaño. El módulo calcula una probabilidad general de que el patrón esté presente basándose en todos estos parámetros y las señales que se estén produciendo e indiquen qué inputs están presentes y cuáles son sus magnitudes. Una forma matemáticamente óptima de conseguir esto la proporciona una técnica llamada modelos ocultos de Márkov. Cuando dichos modelos están organizados según una jerarquía, tal y como pasa en el neocórtex o en los intentos por simular un neocórtex, reciben el nombre de modelos ocultos jerárquicos de Márkov.

Los patrones que se disparan en el neocórtex disparan otros patrones. Patrones parcialmente completos envían señales hacia abajo en la jerarquía conceptual, mientras que patrones completos envían señales hacia arriba en la jerarquía conceptual. Estos patrones neocorticales son el lenguaje del pensamiento. Al igual que el lenguaje, son jerárquicos, pero no son lenguaje en sí mismos. Nuestros

persamientos no son concebidos principalmente mediante los elementos del lenguaje, aunque, debido al hecho de que el lenguaje también existe según jerarquías de patrones en nuestro neocórtex, podemos tener pensamientos basados en el lenguaje. Sin embargo, la mayor parte de los pensamientos vienen representados por estos patrones neocorticales.

Tal y como expuse anteriormente, si fuéramos capaces de detectar las activaciones de los patrones en el neocórtex de alguien, seguiríamos sabiendo muy poco sobre lo que significa la activación de dichos patrones, a no ser que tuviéramos acceso a toda la jerarquía de patrones por encima y por debajo de cada patrón activado. Eso significaría tener acceso a prácticamente todo el neocórtex de dicha persona. Ya nos resulta suficientemente difícil comprender el contenido de nuestros propios pensamientos, pero comprender los de otra persona requieriría dominar un neocórtex diferente al nuestro. Por supuesto, todavía no tenemos acceso al neocórtex de otra persona. En vez de eso, dependemos de sus intentos por expresar sus pensamientos mediante el lenguaje o por otros medios tales como los gestos. La absoluta incapacidad de las personas a la hora de llevar a cabo estas tareas comunicativas añade otra capa más de complejidad, por lo que no es de extrañar que haya tantos malentendidos entre nosotros.

Poseemos dos modos de pensamiento. Uno es el pensamiento indirecto, en el cual los pensamientos se provocan su disparo mutuamente de forma no lógica. Cuando experimentamos una rememoración súbita de un recuerdo de hace años o décadas mientras estamos haciendo algo diferente, como por ejemplo rastrillando hojas o caminando por la calle, la experiencia es recordada (y esto ocurre con todos los recuerdos) como una secuencia de patrones. No visualizamos la escena inmediatamente a no ser que echemos mano de muchos otros recuerdos que nos permitan sintetizar una rememoración más robusta. Si efectivamente visualizamos la escena de ese modo, fundamentalmente estamos creándola en nuestra mente a partir de indicios sobre el momento de la rememoración; el recuerdo en sí no está almacenado bajo la forma de imágenes o visualizaciones. Tal y como comenté anteriormente, los desencadenantes que hicieron que este pensamiento surgiera en nuestra mente pueden ser o no ser evidentes. La secuencia de pensamientos relevantes puede haber sido olvidada inmediatamente. Incluso si la recordamos, se tratará de una secuencia de asociaciones no lineal y enrevesada.

El segundo modo de pensar es el pensamiento directo. Este pensamiento lo usamos cuando intentamos resolver un problema o formular una respuesta organizada. Por ejemplo, podemos estar ensayando en nuestra mente algo que planeamos decirle a alguien, o puede que estemos formulando un pasaje que nos gustaría escribir (a lo mejor en un libro sobre la mente). Cuando pensamos sobre tareas como estas, ya hemos dividido cada una de ellas en una jerarquía de subtareas. Por ejemplo, escribir un libro implica escribir capítulos; cada capítulo

tiene secciones; cada sección tiene párrafos; cada párrafo contiene frases que expresan ideas; cada idea posee elementos que la configuran; cada elemento y cada relación entre elementos es una idea que necesita ser articulada; y así sucesivamente. Al mismo tiempo, nuestras estructuras neocorticales han aprendido ciertas reglas que deben ser respetadas. Si la tarea es la de escribir, entonces deberíamos intentar evitar repeticiones innecesarias; deberíamos intentar asegurarnos de que el lector pueda seguir el hilo de lo que se está escribiendo; deberíamos intentar respetar las reglas de la gramática y del estilo; y así sucesivamente. Por tanto, el escritor tiene que construirse un modelo del lector en su mente, y dicho constructo también es, a su vez, jerárquico. Al llevar a cabo el pensamiento directo, recorremos listas que están en el interior de nuestro neocórtex, cada una de las cuales se expande en amplias jerarquías de sublistas, cada una con sus propias reflexiones. Téngase en cuenta que los elementos de una lista en un patrón neocortical pueden incluir condicionantes, de manera que nuestros pensamientos y acciones subsecuentes dependerán de estimaciones hechas durante el proceso.

Y no solo eso, cada uno de estos pensamientos dirigidos desencadenará jerarquías de pensamientos no dirigidos. Una tormenta permanente de reflexiones acude tanto a nuestras experiencias sensoriales, como a nuestros intentos por realizar pensamietos dirigidos. Nuestra experiencia mental real es compleja y caótica, y está compuesta de estas relampagueantes tormentas de patrones disparados, los cuales cambian más o menos cien veces por segundo.

## *El lenguaje de los sueños*

Los sueños son ejemplos de pensamientos no dirigidos. Hasta cierto punto tienen sentido, ya que el fenómeno en el que un pensamiento desencadena otro está basado en conexiones de patrones reales de nuestro neocórtex. En cuanto a la parte de los sueños que no tiene sentido, intentamos arreglarla mediante nuestra capacidad para fabular. Tal y como describiré en el capítulo 9, los pacientes con el cerebro dividido (cuyo cuerpo calloso, aquello que conecta los dos hemisferios del cerebro, está roto o dañado) fabulan (se inventan) explicaciones con su cerebro izquierdo, el que controla el centro del habla, para explicar lo que el cerebro derecho acaba de hacer con el input al que el cerebro izquiero no ha tenido acceso. Fabulamos todo el tiempo para explicar el resultado de los acontecimientos. Si desea tener un buen ejemplo de esto, simplemente sintonice el resumen diario sobre los movimientos en los mercados financieros. Independientemente de cómo se comporten los mercados, siempre es posible encontrar una buena explicación para lo que ha pasado, y dichas explicaciones a posteriori son muy abundantes (aunque es evidente que si estos

analistas entendieran los mercados realmente no tendrían que desperdiciar su tiempo dando explicaciones).

El acto de fabular también tiene lugar en el neocórtex, por supuesto, ya que se le da muy bien inventarse historias y explicaciones que cumplen ciertos requisitos. Esto lo hacemos cada vez que volvemos a contar una historia. Añadimos detalles que no están a nuestro alcance o que puede que hayamos olvidado de manera que la historia sea más coherente. Esta es la razón por la cual las historias cambian a medida que son contadas una y otra vez por nuevos narradores que tienen diferentes intenciones. Sin embargo, como el lenguaje hablado dio lugar al lenguaje escrito, dimos con una tecnología que podía registrar una versión definitiva de una historia y prevenir de esta manera este tipo de cambios.

El contenido real de un sueño, hasta el punto en que nos acordamos de él, es nuevamente una secuencia de patrones. Estos patrones representan restricciones en una historia, y posteriormente fabulamos una historia que encaja con estas restricciones. La versión del sueño que volvemos a contar (aunque solo sea a nosotros mismos en silencio) es esta fabulación. Cuando volvemos a contar un sueño desencadenamos cascadas de patrones que rellenan el sueño real con respecto a la manera en que lo experimentamos originariamente.

Existe una diferencia fundamental entre los pensamientos de los sueños y nuestro pensar despierto. Una de las lecciones que aprendemos en la vida es que ciertas acciones, incluso pensamientos, no están permitidos en el mundo real. Por ejemplo, aprendemos que no podemos cumplir nuestros deseos de forma inmediata. Existen reglas contra agarrar el dinero de la caja registradora de una tienda, y restricciones a la hora de interactuar con una persona por la que a lo mejor nos sentimos atraídos físicamente. También aprendemos que ciertos pensamientos no están permitidos por que culturalmente están prohibidos. Cuando adquirimos habilidades profesionales, aprendemos las formas de pensar que tienen reconocimiento y son recompensadas en nuestras profesiones, por lo que evitamos patrones de pensamiento que puedan incumplir los métodos y normas de dicha profesión. Muchos de estos tabús son útiles, ya que imponen un orden social y consolidan el progreso. Sin embargo, también pueden evitar el progreso si hacen gala de una ortodoxia inproductiva. Precisamente dicha ortodoxia es la que Einstein superó cuando intentó montarse en un haz de luz mediante sus experimentos mentales.

Las reglas culturales se ejecutan en el neocórtex con la ayuda del cerebro antiguo, especialmente la amígdala. Todo pensamiento que tenemos desencadena otros pensamientos, y algunos de ellos están relacionados con daños asociados. Aprendemos, por ejemplo, que romper una norma cultural, incluso en nuestros pensamientos privados, puede llevarnos al ostracismo, y el neocórtex es consciente de que esto amenaza nuestro bienestar. Si contemplamos dichos pensamientos, la amígdala se dispara y esto genera miedo, lo que generalmente conlleva el abandono

de dicho pensamiento.

Sin embargo, en los sueños estos tabús se relajan, por lo que a menudo soñamos sobre cuestiones que están cultural, sexual o profesionalmente prohibidas. Es como si nuestro cerebro se diera cuenta de que mientras soñamos no somos actores reales. Freud escribió sobre este fenómeno, sin embargo también se dio cuenta de que disfrazamos estos pensamientos peligrosos, por lo menos cuando intentamos rememorarlos, de manera que el cerebro despierto sigue estando protegido contra ellos.

El relajamiento de los tabús profesionales resulta ser útil para resolver problemas de forma creativa. Yo uso todas las noches una técnica mental según la cual pienso sobre un problema en concreto antes de irme a dormir. Esto desencadena secuencias de pensamientos que continúan en mis sueños. Una vez que ya estoy soñando, puedo pensar (*soñar*) sobre soluciones del problema sin la carga que conllevan las restricciones profesionales que me acompañan durante el día. Así, por la mañana puedo acceder a estos pensamientos soñados mientras me encuentro en un estado de duermevela al que a veces se le llama "sueño lúcido" (5).

También son famosos los escritos de Freud sobre la capacidad para tener acceso a la psicología de una persona mediante la interpretación de los sueños. Por supuesto, existe una extensa literatura sobre todos los aspectos de esta teoría, pero la noción fundamental de tener acceso a nosotros mismos por medio del examen de nuestros sueños es algo que tiene sentido. Nuestros sueños son creados por nuestro neocórtex, por lo que su materia puede se reveladora con respecto al contenido y las conexiones que se encuentran en el neocórtex. La relajación de las restricciones de nuestro pensamiento despierto también es útil a la hora de revelar el contenido neocortical al que de otra manera no podríamos acceder de forma directa. También es razonable concluir que los patrones que acaban entrando en nuestros sueños representan cuestiones que nos son importantes, por lo que son pistas para comprender nuestros deseos y miedos no resueltos.

## *Las raíces del modelo*

Tal y como he mencionado anteriormente, en la década de 1980 y 1990 dirigí un equipo que desarrolló la técnica de los modelos jerárquicos ocultos de Márkov para reconocer el habla humana y comprender las expresiones del lenguaje natural. Esta labor fue la predecesora de los sistemas comerciales popularizados hoy en día que reconocen y comprenden lo que estamos tratando de decirles (los sistemas de navegación de coches con los que se puede hablar, el Siri del iPhone, *Google Voice Search* y muchos otros). En el fondo, la técnica que desarrollamos poseía todos los

atributos que describo mediante la PRTM. Incluía una jerarquía de patrones donde cada nivel superior era conceptualmente más abstracto que el inmediatamente inferior. Por ejemplo, en el reconociemiento del habla los niveles incluían patrones básicos de frecuencias sonoras en su nivel más bajo, luego venían los fonemas, después las palabras y luego las frases (que a menudo eran reconocidas como si fueran palabras). Algunos de nuestros sistemas de reconocimiento del habla podían comprender el significado de comandos del lenguaje natural, de manera que niveles todavía más altos incluían estructuras tales como frases con sujeto y predicado. Además, cada módulo de reconocimiento de patrones podía reconocer una secuencia de patrones lineal procedente de un nivel conceptual más bajo. Cada input poseía parámetros en relación a la importancia, el tamaño y la variabilidad del tamaño. Así, existían señales "en sentido descendiente" que indicaban que se estaba a la expectativa de un patrón de nivel más bajo. En el capítulo 7 expongo esta investigación con más detalle.

En 2003 y 2004, el inventor de PalmPilot llamado Jeff Hawkins y Dileep George desarrollaron un modelo cortical jerárquico llamado memoria temporal jerárquica. Junto a la escritora científica Sandra Blakeslee, Hawkins describió este modelo de forma elocuente es su libro *On Intelligence*. Hawkins proporciona un argumento muy sólido a favor de la uniformidad del algoritmo cortical y su organización jerárquica y basada en listas. Existen algunas diferencias importantes entre el modelo presentado en *On Intelligence* y lo que yo expongo en este libro. Tal y como su propio nombre indica, Hawkins enfatiza la naturaleza temporal (basada en el tiempo) de las listas constituyentes. En otras palabras, la dirección de las listas siempre va hacia adelante en el tiempo. Su explicación sobre cómo las características de un patrón bidimensional como por ejemplo la letra impresa "A" tienen una dirección en el tiempo es atribuido al movimiento del ojo. El autor explica que visulizamos las imágenes mediante movimientos sacádicos, movimientos muy rápidos de los ojos de los cuales no somos conscientes. Por tanto, la información que llega al neocórtex no es un conjunto bidimensional de características, sino más bien una lista ordenada temporalmente. Aunque es cierto que nuestros ojos realizan movimientos muy rápidos, la secuencia según la cual inspeccionan las características de un patrón como por ejemplo la letra "A" no siempre tiene lugar en un orden temporal congruente. (Por ejemplo, los movimientos sacádicos no registran en todas las ocasiones el vértice superior de la "A" antes de registrar su cavidad inferior). Además, podemos reconocer un patrón visual que solo esté presente durante unas pocas décimas de milisegundo, que es un periodo de tiempo demasiado corto como para que los movimientos sacádicos puedan escanearlo. Es cierto que los reconocedores de patrones del neocórtex almacenan un patrón en forma de lista y que dicha lista está ordenada, pero el orden no representa necesariamente el tiempo. Amenudo este sí es el caso, pero también puede representar un orden conceptual

espacial o de nivel más alto, tal y como expuse anteriormente.

La diferencia más importante es el conjunto de parámetros que he asignado a cada input en el módulo de reconocimiento de patrones, sobre todo los parámetros del tamaño y de la variabilidad del tamaño. De hecho, en la década de 1980 intentamos reconocer el habla humana sin contar con este tipo de información. El motivo para hacerlo fue que los lingüistas nos decían que la duración de la información no era especialmente importante. Este punto de vista viene ilustrado por los diccionarios en los que se transcribe la pronunciación de cada palabra a modo de cadena de fonemas, por ejemplo en la palabra "steep" como [s] [t] [E] [p], sin indicar la expectativa de duración de cada fonema. De aquí se deduce que si creamos programas para reconocer fonemas y posteriormente nos encontramos con esta secuencia de cuatro fonemas en particular durante una declaración hablada, entonces deberíamos ser capaces de reconocer la pronunciación de dicha palabra. El sistema que construimos utilizando esta estrategia funcionó hasta cierto punto, pero no lo suficientemente bien como para dominar los atributos de un vocabulario extenso, de múltiples interlocutores o de las palabras habladas pronunciadas de forma continua sin hacer pausas. Sin embargo, cuando utilizamos la técnica de los modelos ocultos jerárquicos de Márkov para incorporar la distribución de las magnitudes de cada input, el rendimiento se disparó.

* Para favorecer la claridad del texto hemos dejado los ejemplos que pone el autor sin traducir. *Apple* significa manzana y *pear* significa pera.
* *Pattern Recognition Theory of Mind*. Hemos traducido el término por teoría de la mente basada en el reconocimiento de patrones, sin embargo hemos decido dejar el acrónimo en su forma original.
* Esta palabra tiene muchos significados. Como adjetivo puede significar elevado, inclinado o escarpado; como verbo puede significar remojar o hacer una infusión; y en sentido figurado significa imbuir. Sin embargo, lo importante aquí es su pronunciación: sti:p (con "s" líquida e "i" larga en lugar de "ee").
** Paso.
* *Continuums* en el original.
** Ver la nota a pie de página anterior.
*** Vease la nota a pie de página anterior.
* *Stream of consciousness* en el original.
* Aquí se hace inevitable utilizar un ejemplo extraído del español. El autor utiliza la palabra *goin'* (gerundio del verbo *to go*, que significa ir). La manera correcta de escribir este gerundio sería *going*. Igualmente, en español lo correcto sería escribir parado y no parao.
** Lo mismo que en la nota anterior pero con la palabra demasiado.
*** Véase la nota a pie de página número tres de este capítulo.
* Cima, cumbre, pico o máximo. Se pronuncia pi:k.
* *Linear programming* en el original.

CAPÍTULO CUATRO

# El neocórtex biológico

Como las cosas importantes vienen en una caja, tenemos un cráneo para el cerebro, una funda de plástico para el peine y una cartera para el dinero.

—GEORGE COSTANZA, "THE REVERSE PEEPHOLE", EPISODIO DE LA SERIE DE TELEVISIÓN *SEINFELD*

Ahora, por primera vez, estamos observando el cerebro en funcionamiento de manera global y con tal claridad que deberíamos ser capaces de descubrir los programas generales que están detrás de sus magníficas capacidades.

—J. G. TAYLOR, B. HORWITZ Y K. J. FRISTON

En resumen, la mente trabaja con los datos que recibe de forma muy parecida a como un escultor trabaja con el bloque de piedra. En cierto sentido, la estatua había estado ahí toda la eternidad. Sin embargo, a su lado se encontraban miles de estatuas diferentes y solo gracias al escultor pudo la elegida ser liberada del resto. De igual manera, el mundo de cada uno de nosotros, independientemente de lo diferentes que sean nuestras opiniones sobre él, se encuentra sumergido en el caos primordial de las sensaciones, lo cual provocó que la mera *materia* fuera pensada por cada uno de nosotros de forma diferente. Si se prefiere, podemos decir que mediante nuestra capacidad de razonamiento somos capaces de descomponer las cosas hasta devolverlas a aquella oscura y deslavazada continuidad de espacio y nubes en movimiento compuestas por enjambres de átomos que la ciencia tilda de único mundo verdadero. Sin embargo, el mundo que *sentimos y en el que vivimos* sigue siendo el mundo que nuestros ancestros y nosotros, a golpe de lentas y acumulativas decisiones, hemos liberado del resto de mundos posibles. Lo hemos conseguido, tal y como hace el escultor, mediante el simple rechazo de ciertas partes de las cosas que se nos presentan. ¡A diferentes escultores, diferentes estatuas hechas de la misma piedra! ¡A diferentes mentes, diferentes mundos nacidos del mismo monótono e inexpresivo caos! Mi mundo no es sino uno entre un millón de otros mundos igualmente sumergidos e igualmente reales a aquellos de los que se ha separado. ¡Qué diferente debe ser el mundo en la consciencia de la hormiga, de la sepia o del cangrejo!"

—WILLIAM JAMES

¿Es la inteligencia el objetivo, o tan siquiera *uno* de los objetivos, de la evolución biológica? Steven Pinker escribe: "somos unos chovinistas sobre nuestro cerebro al considerarle como el objetivo de la evolución" (1) y continúa diciendo que "eso no tiene sentido […]. La selección natural no aspira a nada que sea remotamente parecido a la inteligencia. El proceso es dirigido mediante diferencias en el índice de supervivencia y reproducción de organismos replicantes pertenecientes a un medio concreto. Con el tiempo, los organismos adquieren diseños que les hacen adaptarse para que puedan sobrevivir y reproducirse en el medio o periodo en el que estén, nada les empuja en una dirección

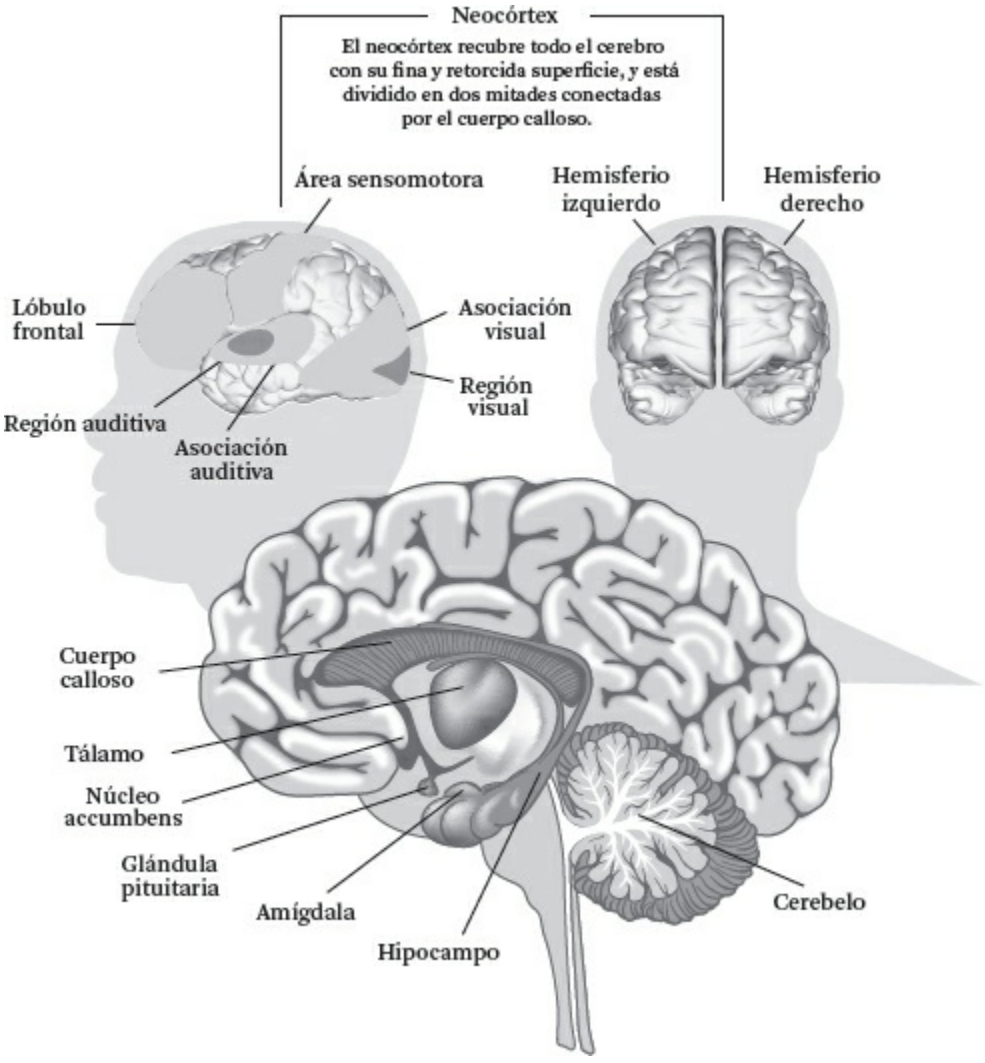
concreta a no ser el éxito que alcancen en dicho medio y periodo". Así, Pinker llega a la conclusión de que "la vida es un arbusto de denso ramaje, no una escala o escalera, y los organismos vivos se encuentran en las puntas de las ramas, no en peldaños inferiores".

En lo que respecta al cerebro humano, se cuestiona si los "beneficios son superiores a los costes". Entre los costes señala que "el cerebro es voluminoso. La pelvis femenina apenas sí se amolda a la cabeza excesivamente grande de un bebé. Este diseño mata a muchas mujeres durante el parto y necesita de una manera de andar basculante que hace que las mujeres sean biomecánicamente menos eficientes a la hora de andar que los hombres. Asimismo, una cabeza grande banboleándose en el cuello nos hace más vulnerables a sufrir lesiones mortales en un accidente, como por ejemplo una caída". Pinker procede a enumerar otras deficiencias, tales como el consumo energético del cerebro, su lento tiempo de reacción y el largo proceso del aprendizaje.

Aunque en su superficie todas estas afirmaciones son certeras (pese a que muchas de mis amigas andan mejor que yo), a Pinker se le escapa la cuestión principal. Es cierto que biológicamente la evolución no va en una dirección específica. La evolución es un método de búsqueda que ciertamente rellena por completo el "denso ramaje" de la naturaleza. También es cierto que los cambios evolutivos no tienen que desplazarse *necesariamente* en la dirección de una mayor inteligencia, sino que se desplazan en *todas* las direcciones. Existen muchos ejemplos de criaturas exitosas que se han mantenido relativamente invariables durante millones de años. (Los caimanes, por ejemplo, datan de hace 200 millones de años y muchos microorganismos se remontan a todavía mucho antes). Sin embargo, en el proceso de rellenado de la infinidad de ramas evolutivas, una de las direcciones *sí* que se desplaza hacia una mayor inteligencia. Esto es lo relevante para nuestro debate.

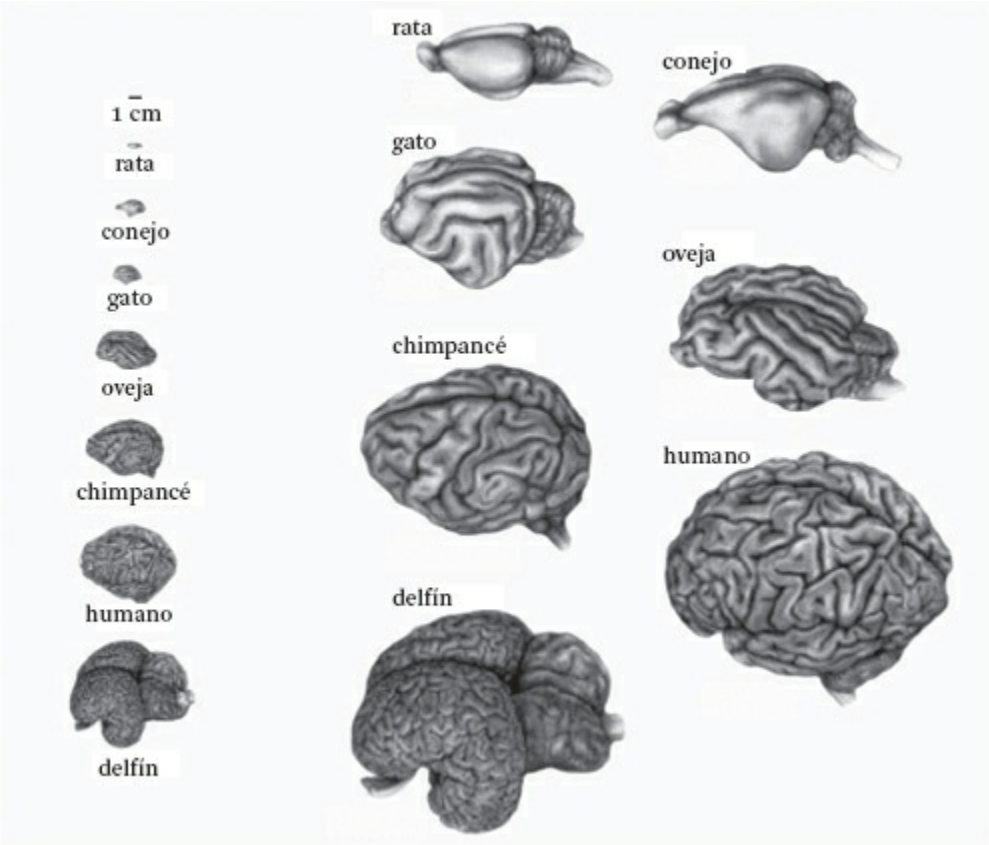
Supongamos que tenemos un gas azul en un frasco. Cuando quitamos la tapadera, no hay ningún mensaje que llegue a todas las moléculas del gas diciendo: "oíd, chicos, han quitado la tapadera del frasco, vayamos hacia la abertura de arriba y hacia la libertad". Las moléculas se limitan a hacer lo que siempre hacen, moverse hacia todos los lados sin dirección aparente. No obstante, algunas moléculas cerca de la parte de arriba sí que abandonarán el frasco, y con el tiempo la mayoría seguirá su ejemplo. Cuando la evolución biológica se tropezó con un mecanismo neuronal capaz de aprender jerárquicamente, descubrió que era inmensamente útil para el objetivo de la evolución, es decir, para la supervivencia. Además, los beneficios de tener un neocórtex se acentuaban cuando circunstancias rápidamente cambiantes empujaban hacia un rápido aprendizaje. Con el tiempo, especies de todos los tipos (tanto plantas, como animales) pueden aprender a adaptarse a circunstancias

cambiantes, pero sin un neocórtex se ven obligadas a utilizar el proceso de la evolución genética. Una especie sin neocórtex puede tardar una enorme cantidad de generaciones (miles de años) en aprender nuevos comportamientos de importancia (o, en el caso de las plantas, nuevas estrategias adaptativas). La ventaja fundamental que conllevó el neocórtex en lo que se refiere a la supervivencia fue que podía aprender en cuestión de días. Si una especie se encuentra en circunstancias que han cambiado drásticamente y un miembro de dicha especie inventa, descubre o simplemente se topa con una manera de adaptarse a dicho cambio (los tres métodos son diferentes variantes de la innovación), otros individuos se darán cuenta de ello, procediendo a aprender y copiar dicho método, que rápidamente se esparcirá como una plaga por toda la población. Así, hace unos 65 millones de años, la extinción masiva del Cretácico-Terciario produjo la rápida desaparición de muchas especies no dotadas de neocórtex que no pudieron adaptarse lo suficientemente rápido a un medio súbitamente alterado. Esto marcó el punto de inflexión para que los mamíferos capaces de albergar un neocórtex invadieran su nicho ecológico. De esta manera, la evolución biológica descubrió lo valioso que era el aprendizaje jerárquico del neocórtex, y esta región del cerebro continuó creciendo en tamaño hasta que prácticamente ocupó todo el cerebro del *homo sapiens*.



**Distribución física de las regiones fundamentales del cerebro.**

Descubrimientos hechos en neurociencia han demostrado fehacientemente el papel fundamental jugado por las capacidades jerárquicas del neocórtex. Asimismo, estos descubrimientos nos han proporcionado evidencias que respaldan la teoría de la mente basada en el reconocimiento de patrones (PRTM). Estas evidencias están distribuidas a lo largo de muchas observaciones y análisis, parte de los cuales revisaré en este libro.




**El neocórtex de diferentes mamíferos.**

El psicólogo canadiense Donald O. Hebb (1904–1985) realizó un primer intento por explicar las bases neurológicas del aprendizaje. En 1949 describió un mecanismo según el cual las neuronas cambian fisiológicamente según sus experiencias, con lo que sentó las bases de lo que se conoce como plasticidad en el aprendizaje y en el cerebro: "Asumamos que la persistencia o repetición de una actividad reverberatoria (o 'traza') tiende a inducir cambios celulares permanentes que ayudan a su estabilidad.[…] Cuando un axón de la célula A está lo suficientemente cercano a excitar una célula B, y repetitiva o persistentemente toma parte en su disparo, tiene lugar un proceso de crecimiento o cambio metabólico en una o ambas células de manera que la eficiencia de A, siendo esta una de las células que hace que B se dispare, aumenta" (2). Esta teoría ha sido resumida diciendo que "las células que se disparan juntas, permanecerán conectadas" y se ha dado a conocer bajo el nombre del aprendizaje de Hebb. Hay aspectos de la teoría hebbiana

que han sido confirmados, ya que está claro que el ensamblado del cerebro puede crear nuevas conexiones y reforzarlas basándose en la actividad de estas conexiones. De hecho, ya podemos ver en escáneres cerebrales cómo las neuronas desarrollan tales conexiones en el cerebro. Así, las "redes neuronales" artificiales se basan en el modelo de Hebb para el aprendizaje neuronal.

La tesis central de la teoría hebianna es que la unidad básica del aprendizaje en el neocórtex es la neurona. La teoría de la mente según el reconocimiento de patrones que expongo en este libro se basa en una unidad fundamental diferente: no la neurona en sí, sino más bien un conjunto de neuronas que yo calculo que ascienden a alrededor de cien. La unión y potencia sinápticas *dentro* de cada unidad son relativamente estables y vienen determinadas genéticamente, es decir, que la organización dentro de cada módulo de reconocimiento de patrones viene determinada por el diseño genético. El aprendizaje tiene lugar en la creación de conexiones *entre* estas unidades, no dentro de ellas, y probablemente en las potencias sinápticas de dichas conexiones entre unidades.

Un respaldo reciente al hecho de que el módulo básico del aprendizaje sea un módulo compuesto de docenas de neuronas proviene del neurocientífico suizo Henry Markram (nacido en 1962), cuyo ambicioso Blue Brain Project para simular el cerebro humano en su conjunto viene descrito en el capítulo 7 de este libro. En un trabajo de investigación de 2011, Markram describe cómo mientras escaneaba y analizaba neuronas reales del neocórtex de mamíferos, lo que hacía realmente era "buscar evidencias de uniones hebbianas al nivel más elemental del córtex". Escribe que, en lugar de eso, lo que encontró fueron "uniones imprecisas [cuyas] conectividades y potencias sinápticas son altamente predecibles y limitadas". Así, llega a la conclusión de que "estos descubrimientos implican que la experiencia no puede moldear fácilmente las conexiones sinápticas de estas uniones" y especula que "sirven como pilares fundamentales innatos parecidos a piezas de *Lego* destinados a la percepción y que la adquisición de recuerdos implica la combinación de estos pilares fundamentales para formar constructos complejos". Continúa diciendo:

> Los conjuntos neuronales se conocen desde hace décadas, […] pero faltaban evidencias directas de cúmulos de neuronas conectados sinápticamente. […] Dado que todos estos conjuntos son similares en cuanto a su topología y pesos sinápticos, y no vienen modelizados por ninguna experiencia específica, los consideramos como conjuntos innatos. […] La experiencia desempeña solamente un papel menor en la determinación de las conexiones y pesos sinápticos dentro de estos conjuntos.[…] Nuestra investigación encontró evidencias [de] innatos conjuntos similares al Lego contenidos en una cuantas docenas de neuronas. […] Las conexiones entre conjuntos pueden hacer que se combinen. Así, los conjuntos se convierten en superconjuntos ubicados dentro una capa neocortical; después en conjuntos de orden superior dentro de una columna cortical; después en conjuntos de orden todavía mayor dentro de una región del cerebro; y finalmente en el conjunto de orden más alto posible que viene representado por el cerebro al completo.[…] La adquisición de recuerdos es muy similar a construir un Lego. Cada conjunto equivale a un bloque de Lego portador de alguna pieza de conocimiento innato elemental sobre cómo procesar, percibir y responder ante el mundo. […] Cuando bloques diferentes se juntan, forman una combinación única de estos apercibimientos innatos que representa una experiencia y conocimiento

específico del individuo” (3).

Los “bloques de Lego” que propone Markram son perfectamente congruentes con los módulos de reconocimiento de patrones que he descrito. En una correspondencia por email Markram describe estos “bloques de Lego” como “contenido compartido y conocimiento innato” (4). Yo añadiría que el propósito de estos módulos es reconocer patrones, recordarlos y predecirlos basándose en patrones parciales. Téngase en cuenta que la estimación que hace Markram de que cada módulo contiene “varias docenas de neuronas” está basada solamente en la capa V del neocórtex. Efectivamente, la capa V es rica en neuronas, pero si nos basamos en la proporción habitual de neuronas que encontramos en las seis capas esto se traduciría en un orden de magnitud de alrededor de 100 neuronas por módulo, cifra que es congruente con mis estimaciones.

El cableado y aparente modularidad del neocórtex son facetas conocidas desde hace años, pero esta investigación ha sido la primera en demostrar la estabilidad de estos módulos a medida que el cerebro realiza sus procesos dinámicos.

Otra investigación reciente, realizada en el *Massachusetts General Hospital*, financiada por los *National Institutes of Health* y la *National Science Foundation* y publicada en la edición de marzo de 2012 de la revista *Science*, también demuestra la estructura regular de las conexiones a lo largo del neocórtex (5). El artículo describe el cableado del neocórtex según un patrón en red similar al orden que siguen las calles de una ciudad bien urbanizada: “en lo fundamental, la estructura general del cerebro acaba por parecerse a Manhattan, donde nos encontramos con un plano de dos dimensiones y con un tercer eje, un ascensor que se mueve en la tercera dimensión”, escribe Van J. Wedeen, el neurocientífico y físico de Harvard que dirigió la investigación.

En un *podcast* de la revista *Science*, Wedeen describe el significado de la investigación: “esta fue una investigación sobre la estructura tridimensional de los senderos del cerebro. Aunque los científicos ya llevaban más o menos un siglo reflexionando sobre los senderos del cerebro, la imagen o modelo típico que nos viene a la mente es que estos senderos se parecen a una fuente de espaguetis, senderos separados cuya relación entre ellos sigue un patrón espacial muy poco definido. Mediante la imagen por resonancia magnética*, fuimos capaces de investigar esta cuestión de forma experimental. Lo que descubrimos es que, en vez de senderos independientes o caprichosamente organizados, se trata más bien de un conjunto de senderos cerebrales que tomados en su totalidad conforman una única estructura extremadamente simple. Básicamente, tienen aspecto de cubo. En lo fundamental funcionan según tres direcciones perpendiculares. En cada una de estas tres direcciones los senderos son enormemente paralelos entre ellos y están organizados matricialmente. Así, en vez de espaguetis independientes, constatamos

que la conectividad del cerebro es, en cierto sentido, una única estructura coherente".

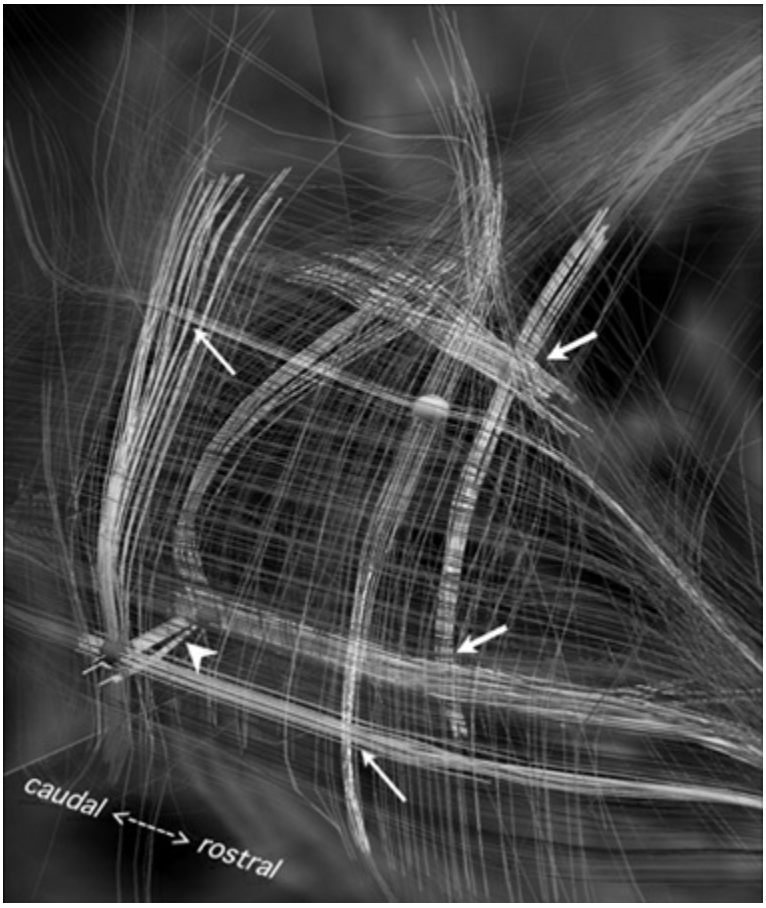
Al igual que la investigación de Markram muestra cómo los módulos de neuronas se repiten a lo largo del neocórtex, la investigación de Wedeen muestra un patrón extraordinariamente ordenado en las conexiones entre módulos. El cerebro parte de un gran número de "conexiones a la espera" a las que los módulos de reconocimiento de patrones pueden conectarse. Por tanto, si un módulo en concreto desea conectarse con otro, no tiene que generar un axón en uno y una dendrita en el otro para salvar la distancia física que dista entre ellos. Simplemente puede emplear una de estas conexiones axonales a la espera y conectarse a los extremos de la fibra. Tal y como Wedeen y sus colegas escriben, "los senderos del cerebro siguen un plan básico definido por […] la embriogénesis temprana. Así, los senderos de un cerebro maduro presentan un imagen de estos tres gradientes primordiales, físicamente deformados por el desarrollo". En otras palabras, a medida que aprendemos y tenemos experiencias, los módulos de reconocimiento de patrones del neocórtex se van conectando con estas conexiones preestablecidas que fueron creadas cuando todavía éramos embriones.

Existe un tipo de chip electrónico llamado FPGA* que se basa en un principio parecido. Este chip contiene millones de módulos que llevan a cabo funciones lógicas junto con las conexiones en espera. Al ser utilizadas, estas conexiones se activan o desactivan mediante señales electrónicas que les permiten poner en práctica alguna de sus capacidades.

En el neocórtex, aquellas conexiones de larga distancia que no son utilizadas son, eventualmente, podadas. Esta es una de las razones por las cuales la adaptación de una región cercana del neocórtex para compensar el daño sufrido en otra región no resulta ser algo tan efectivo como usar la región original. Según la investigación de Wedeen, las conexiones primigenias son extraordinariamente ordenadas y repetitivas, igual que los propios módulos, y su patrón en red se utiliza a modo de "guía de conectividad" en el neocórtex. Este patrón ha sido encontrado en todos los primates y cerebros humanos que han sido estudiados. Además, se hace patente por todo el neocórtex, partiendo de las regiones que se encargan de los patrones sensoriales básicos hasta llegar a aquellas que tienen que ver con las emociones de más alto nivel. El artículo de Wedeen en la revista *Science* llega a la conclusión de que la "estructura en red de los senderos cerebrales es omnipresente, coherente y continua con respecto a los tres ejes principales del desarrollo". De nuevo, esto hace referencia a un algoritmo común a todas las funciones neocorticales.

Desde hace mucho se sabe que por lo menos ciertas regiones del neocórtex son jerárquicas. La región mejor estudiada es el córtex visual, que está dividido en áreas conocidas como V1, V2 y MT (también llamada V5). A medida que se avanza hacia áreas más altas de esta región ("más altas" en el sentido de procesamiento

conceptual, no en sentido físico, ya que el neocórtex siempre tiene el grosor de un reconocedor de patrones), las propiedades que pueden ser reconocidas se vuelven más abstractas. V1 reconoce perfiles muy básicos y formas primitivas. V2 puede reconocer contornos, la disparidad de imágenes surgida de cada uno de los ojos, la orientación espacial y si un trozo de imagen forma parte de un objeto o de un segundo plano (6). Las regiones de más alto nivel del neocórtex reconocen conceptos tales como la identidad de los objetos y caras, así como sus movimientos. También se sabe desde hace tiempo que la comunicación a través de esta jerarquía se produce tanto en sentido ascendente como descendente, y que las señales pueden ser tanto excitativas como inhibidoras. Tomaso Poggio, neurocientífico del MIT nacido en 1947, ha estudiado concienzudamente la visión en el cerebro humano, y su labor investigadora durante los últimos treintaicinco años ha sido decisiva a la hora de fundamentar el aprendizaje jerárquico y el reconocimiento de patrones en los niveles "tempranos" (conceptualmente más bajos) del neocórtex visual (7).




**La estructura en red altamente regular de las primeras conexiones del neocórtex demostrada por la investigación de los *National Institutes of Health*.**

**Otra perspectiva de la estructura en red regular de las conexiones neocorticales.**

**La estructura en red descubierta en el neocórtex es notablemente parecida a lo que se ha dado en llamar conmutación de barras cruzadas, que se utiliza es los circuitos integrados y en las placas de circuitos.**

Nuestra comprensión de los niveles jerárquicos más bajos del neocórtex visual es coherente con la PRTM que describí en el capítulo anterior. Asimismo, la monitorización de la naturaleza jerárquica del procesamiento neocortical se ha

extendido recientemente mucho más allá de estos niveles. J. Felleman, profesor de neurobiología de la Universidad de Texas, y sus colegas han delineado la "organización jerárquica del córtex cerebral [de] 25 áreas neocorticales", que incluyen áreas tanto visuales como áreas de nivel más alto que combinan patrones pertenecientes a múltiples sentidos. Lo que descubrieron a medida que ascendían por la jerarquía neocortical fue que el procesamiento de los patrones se volvía más abstracto, constaba de áreas espaciales más grandes y conllevaba periodos de tiempo más largos. Además, en cada conexión encontraron comunicaciones dentro la jerarquía que eran tanto ascendentes como descendentes (8).

Investigaciones recientes nos permiten ampliar en gran medida estas observaciones hasta regiones que están mucho más allá del córtex visual y que llegan incluso a áreas asociativas, aquellas que combinan inputs procedentes de múltiples sentidos. Una investigación publicada en 2008 por el profesor de psicología de Princeton Uri Hasson y sus colegas demuestra que los fenómenos observados en el córtex visual tienen lugar a lo largo de una gran variedad de áreas neocorticales: "es bien sabido que las neuronas que se encuentran en los senderos corticales visuales poseen campos receptivos cuya extensión va en aumento. He aquí un principio organizativo básico del sistema visual. […] Los hechos del mundo real acaecen no solo sobre amplias regiones espaciales, sino que también lo hacen durante extensos periodos de tiempo. Por tanto, nuestra hipótesis es que una jerarquía análoga a aquella descubierta en los tamaños de los campos receptivos del espacio también debe de existir en el caso de las características de respuesta temporal de diferentes regiones del cerebro". Esto es exactamente lo que descubrieron, lo cual les permitió concluir que "de forma similar a la ya conocida jerarquía cortical de los campos receptivos espaciales, existe una jerarquía en el cerebro humano compuesta de ventanas receptivas temporales cuya duración va progresivamente en aumento" (9).

El argumento más poderoso a favor de la universalidad de procesamiento en el neocórtex es la generalizada evidencia de su plasticidad (no solo en cuanto al aprendizaje, sino también en lo que respecta a su intercambiabilidad). En otras palabras, una región es capaz de realizar el trabajo de otras regiones, lo cual implica la existencia de un algoritmo común a lo largo de todo el neocórtex. Así, una gran cantidad de investigación en el campo de la neurociencia se ha centrado en identificar qué regiones del neocórtex son las responsables de los diferentes tipos de patrones. La técnica clásica utilizada para determinar esto ha sido el aprovechamiento del daño cerebral producido por una lesión o derrame para relacionar la pérdida de funcionalidad con las regiones específicamente dañadas. Por ejemplo, cuando percibimos que de repente alguien que acaba de sufrir un daño en la región giro fusiforme tiene dificultades a la hora de reconocer caras, pero sin embargo sigue siendo capaz de identificar a personas por sus patrones de voz y de lenguaje, podemos construir la hipótesis de que dicha región tiene algo que ver con

el reconocimiento facial. La suposición adyacente a la que se ha llegado es que cada una de estas regiones está diseñada para reconocer y procesar un tipo particular de patrón. Por tanto, regiones físicas concretas han sido asociadas a tipos de patrones concretos, ya que en circunstancias normales es así como la información fluye. Sin embargo, cuando el flujo normal de información se ve interrumpido por alguna causa, otra región del neocórtex puede entrar en acción y sustituir a la región original.

La plasticidad ha sido ampliamente observada por los neurólogos, que se dieron cuenta de que pacientes con un daño cerebral producido por una lesión o derrame pueden volver a aprender las mismas capacidades en otro área del neocórtex. Quizás el ejemplo más drástico de plasticidad sea una investigación del año 2011 llevada a cabo por la neurocientífica norteamericana Marina Bedny y sus colegas sobre lo que le ocurre al córtex visual de personas con ceguera congénita. La creencia popular era que las primeras capas del córtex visual tales como la V1 y la V2 estaban intrínsecamente relacionadas con patrones de muy bajo nivel (por ejemplo bordes y curvas), mientras que el córtex frontal (la región evolutivamente nueva del córtex que poseemos en nuestras exclusivas frentes de grandes dimensiones) estaba intrínsecamente relacionado con los patrones más complejos y sutiles correspondientes al lenguaje y otros conceptos abstractos. No obstante, Bedny y sus colegas descubrieron que "los humanos parecen haber desarrollado regiones cerebrales en el córtex temporal y central izquierdo que únicamente son capaces de procesar el lenguaje. Sin embargo, los individuos congénitamente ciegos también activan el córtex visual durante algunas tareas verbales. Proporcionamos evidencias de que, de hecho, esta actividad del córtex visual refleja el procesamiento del lenguaje. Así, hemos descubierto que en individuos congénitamente ciegos el córtex visual izquierdo se comporta de forma similar a las clásicas regiones del lenguaje. […] Nuestra conclusión es que las regiones del cerebro que se piensa que se han desarrollado para la visión pueden hacerse cargo del procesamiento del lenguaje como resultado de una experiencia temprana" (10).

Consideremos las implicaciones de esta investigación. Esta investigación significa que las regiones neocorticales que se encuentran relativamente alejadas las unas de las otras y que conceptualmente también han sido consideradas como muy diferentes (indicaciones visuales primitivas en contraposición a conceptos abstractos del lenguaje) utilizan esencialmente el mismo algoritmo. Por tanto, las regiones que procesan estos tipos de patrones tan dispares pueden sustituirse las unas por las otras.

El neurocientífico Daniel E. Feldman, de la *University of California at Berkeley*, escribió en el año 2009 un detallado análisis de lo que llamó "mecanismos sinápticos de la plasticidad del neocórtex" y encontró evidencias de este tipo de

plasticidad por todo el neocórtex. Escribe que “la plasticidad permite que el cerebro aprenda y recuerde patrones pertenecientes al mundo sensorial, que perfeccione los movimientos […] y que recupere su funcionamiento después de una lesión”. Además, añade que esta plasticidad es posible gracias a “cambios estructurales que incluyen la formación, la eliminación y el rediseño morfológico de las sinapsis corticales y de las espinas dendríticas” (11).

Otro sorprendente ejemplo de la plasticidad neocortical (y por tanto de la uniformidad del algoritmo neocortical) ha sido expuesto recientemente por científicos de la *University of California at Berkeley*. Conectaron chips de microelectrodos previamente implantados para que recogieran las señales provenientes de una región específica de la corteza motora de los ratones que controla el movimiento de sus bigotes. Dispusieron su experimento de forma que los ratones recibieran una recompensa si habían controlado estas neuronas para que se disparasen según un patrón mental determinado que no fuera el movimiento de sus bigotes. El patrón necesario para recibir la recompensa involucraba una tarea mental que sus neuronas frontales normalmente no realizan. No obstante, los ratones pudieron realizar esta hazaña mental fundamentalmente pensando con sus neuronas motoras a la vez que las escindían del control de sus movimientos motores (12). La conclusión es que la corteza motora, la región del neocórtex responsable de coordinar el movimiento muscular, también utiliza el algoritmo neocortical estándar.

Sin embargo, existen varias razones por las que una capacidad o un área de conocimiento que se ha vuelto a aprender utilizando una nueva área del neocórtex para remplazar aquella área que ha sido dañada no será necesariamente tan buena como la original. Primero, porque fue necesaria toda una vida para aprender y perfeccionar una capacidad en particular. Volver a aprenderla en otra área del neocórtex no generará inmediatamente los mismos resultados. Más importante todavía es que dicha nueva área del neocórtex no se ha limitado a esperar sin hacer nada hasta que una región resultara dañada. Esa área también ha desarrollado funciones vitales y por tanto se mostrará renuente a la hora de abandonar sus patrones neocorticales para compensar el daño sufrido por la otra región. Por ejemplo, puede comenzar por degradar de forma sutil sus capacidades adquiridas y no liberar tanta cantidad de espacio cortical como el que originalmente utilizaron las capacidades que ahora tienen que volver a ser aprendidas.

Hay una tercera razón por la cual la plasticidad tiene sus límites. Dado que en la mayor parte de las personas ciertos tipos de patrones fluyen a través de regiones específicas (como por ejemplo las caras que son procesadas, que pasan por el giro fusiforme), estas regiones han sido optimizadas mediante la evolución biológica para dichos tipos de patrones. Tal y como expongo en el capítulo 7, se observa el mismo hecho en nuestros desarrollos neocorticales digitales. Podríamos reconocer el habla mediante nuestros sistemas de reconocimiento de caracteres y viceversa, pero los

sistemas del habla fueron optimizados para el habla y, de igual forma, los sistemas de reconocimiento de caracteres fueron optimizados para los caracteres escritos, de manera que el rendimiento se vería en cierta manera reducido si sustituyéramos uno por otro. De hecho, hemos utilizado algoritmos evolutivos (genéticos) para conseguir esta optimización, lo que resulta en una simulación de lo que la biología hace de forma natural. Dado que las caras han estado fluyendo por el giro fusiforme de la mayor parte de las personas durante cientos de miles de años (o más), la evolución biológica ha tenido tiempo de desarrollar una capacidad que favorece el procesamiento de dichos patrones en dicha región. Para ello, utiliza el mismo algoritmo básico, pero lo orienta hacia las caras. Tal y como el neurocientífico holandés Randal Koene escribió, "el [neo]córtex es muy uniforme, cada columna o minicolumna puede, en principio, realizar lo que las demás" (13).

Exhaustivas investigaciones recientes apoyan la observación de que los módulos de reconocimiento de patrones se autoorganizan según los patrones a los que están expuestos. Por ejemplo, la neurocientífica Yi Zuo y sus colegas observaron cómo, al mismo tiempo que nuevas "espinas dendríticas" formaban conexiones entre células nerviosas, los ratones aprendían una nueva capacidad (en este caso, agarrar una semilla a través de una ranura) (14).

Investigadores del *Salk Institute* han descubierto que aparentemente esta autoorganización fundamental de los módulos del neocórtex está controlada por solo un puñado de genes. Estos genes y este método de autoorganización también son uniformes a lo largo del neocórtex (15).

Muchas otras investigaciones documentan estos atributos del neocórtex; sin embargo, hagamos un resumen de lo que podemos observar partiendo de la literatura neurocientífica y de nuestros propios experimentos mentales.

La unidad básica del neocórtex es un módulo de unas cien neuronas aproximadamente. Estos módulos están entrelazados los unos con los otros en el interior de cada columna neocortical, de manera que cada módulo no es claramente diferenciable. El patrón de las conexiones y la potencia sináptica dentro de cada módulo son relativamente estables. Así, son las conexiones y las potencias sinápticas *entre* módulos las que representan el aprendizaje.

Hay alrededor de mil billones ($10^{15}$) de conexiones en el neocórtex, pero sin embargo en el genoma solo hay 25 millones de bytes de información concernientes al diseño (y esto después de haber realizado una compresión en la que no se producen pérdidas) (16). Por tanto, las propias conexiones no pueden venir predeterminadas genéticamente. Es posible que parte de este aprendizaje sea el resultado de la interpelación del neocórtex al cerebro antiguo, pero aun así eso solo representaría una cantidad de información relativamente pequeña. En conjunto, las conexiones entre módulos se crean a partir de la experiencia (educación en lugar de naturaleza).

El cerebro no posee la flexibilidad suficiente como para que cada módulo de reconocimiento de patrones neocortical pueda simplemente conectarse a cualquier otro módulo, a diferencia de lo que fácilmente podemos programar en nuestros ordenadores o en la web. En el cerebro tiene que formarse una conexión física compuesta por un axón que se conecta a una dendrita. Además, todos y cada uno de nosotros empezamos la vida con unas reservas de posibles conexiones neuronales muy amplias. Tal y como demuestra la investigación de Wedeen, estas conexiones se organizan de una manera muy repetitiva y ordenada. La conexión terminal a estos axones a la espera tiene lugar según los patrones que cada reconocedor de patrones neocortical haya reconocido. Las conexiones no usadas acaban por ser podadas. Además, estas conexiones se construyen jerárquicamente, lo cual refleja el orden jerárquico natural de la realidad. He aquí el punto fuerte del neocórtex.

El algoritmo básico de los módulos de reconocimiento de patrones neocorticales es el mismo en todo el neocórtex, desde los módulos de "nivel bajo" que tienen que ver con los patrones sensoriales más básicos, hasta los módulos de "nivel alto" que reconocen los conceptos más abstractos. La amplia evidencia de la plasticidad e intercambiabilidad de las regiones neocorticales da testimonio de esta importante observación. No obstante, existe cierta optimización de las regiones que tienen que ver con tipos concretos de patrones, pero esto es un efecto de segundo orden: el algoritmo fundamental es universal.

Las señales suben y bajan por la jerarquía conceptual. Una señal que sube significa: "he detectado un patrón". Una señal que baja significa: "estoy a la espera de que se presente tu patrón" y básicamente se puede considerar como una predicción. Tanto las señales hacia arriba como hacia abajo pueden ser o bien excitativas o bien inhibidoras.

Cada patrón tiene su propio orden y este no es fácilmente reversible. Incluso los patrones que parecen tener aspectos multidimensionales vienen representados por una secuencia unidimensional de patrones de más bajo nivel. Además, un patrón es una secuencia ordenada de otros patrones, de manera que cada reconocedor es intrínsecamente recursivo. Así, puede haber muchos niveles de jerarquías.

Existe una gran cantidad de redundancia en los patrones que aprendemos, especialmente en los importantes. El reconocimiento de patrones, como por ejemplo el reconocimiento de objetos comunes o caras, utiliza el mismo mecanismo que nuestros recuerdos, que no son más que patrones que hemos aprendido. Asimismo, se encuentran almacenados como secuencias de patrones (básicamente son historias). Dicho mecanismo también es usado para aprender y llevar a cabo movimientos físicos en el mundo. La redundancia de los patrones es lo que nos permite reconocer objetos, personas e ideas, incluso cuando estos presentan variaciones y se presentan en contextos diferentes. Los parámetros de tamaño y de variabilidad del tamaño

también permiten que el neocórtex codifique la variación de la magnitud con respecto a dimensiones diferentes (por ejemplo, la duración en el caso del sonido). Una manera en la que estos parámetros de magnitud pueden ser codificados es simplemente por medio de patrones múltiples con números diferentes de inputs repetidos. Por ejemplo, podría haber patrones para la palabra hablada "steep" con números diferentes de [E], la vocal larga repetida, lo cual indicaría que la repetición de la [E] es variable. Esta estrategia no es matemáticamente equivalente a tener los parámetros de tamaño explícitos y en la práctica no funciona igual de bien en absoluto, pero se trata de una estrategia para codificar la magnitud. La evidencia más fuerte que tenemos de estos parámetros es que estos son necesarios en nuestros sistemas de inteligencia artificial para lograr niveles de precisión que se aproximen a los niveles humanos.

El resumen anterior pone de manifiesto las conclusiones que podemos sacar a partir de las muestras pertenecientes a los resultados de las investigaciones que he expuesto más arriba, así como de los experimentos mentales a los que he hecho referencia anteriormente. Así, mantengo que el modelo que he presentado es el único que satisface todas las condiciones que la investigación y nuestros experimentos mentales nos imponen.

Para terminar, existe un ejemplo más que corrobora lo anterior. Las técnicas que hemos desarrollado durante las últimas décadas en el campo de la inteligencia artificial para reconocer y procesar inteligentemente fénomenos del mundo real (como por ejemplo el habla humana y el lenguaje escrito) y para comprender documentos en lenguaje natural resultan ser matemáticamente similares al modelo que he expuesto más arriba. Por tanto, también ellos son ejemplos de la PRTM. El campo de la inteligencia artificial no ha intentado copiar de forma explícita el cerebro, pero sin embargo ha desembocado en técnicas que son esencialmente equivalentes.

* *Magnetic resonance imaging* en el original.
* *Field Programmable Gate Array*.

CAPÍTULO CINCO

# El cerebro antiguo

Tengo un cerebro antiguo, pero una gran memoria.

—AL LEWIS

Henos aquí, en medio de este nuevo mundo, con nuestro primitivo cerebro adaptado para la simple vida en la cueva y con tremendas fuerzas a nuestra disposición que somos lo suficientemente listos como para haber liberado. Sin embargo, las consecuencias de esto no las podemos comprender.

—ALBERT SZENT-GYÖRGYI

Nuestro cerebro antiguo, el que tenemos desde antes de ser mamíferos, no ha desaparecido. De hecho, todavía nos provee de gran parte de nuestra motivación para buscar gratificaciones y evitar peligros. Sin embargo, estos objetivos están modulados por nuestro neocórtex, que es lo que domina el cerebro humano tanto en masa como en actividad.

Los animales solían vivir y sobrevivir sin neocórtex, y de hecho todos los animales no mamíferos siguen haciéndolo a día de hoy. Así, podemos ver el neocórtex humano como el gran sublimador, ya que nuestra motivación primitiva por evitar a un depredador grande hoy puede ser transformada por el neocórtex en la finalización de un encargo que impresione a nuestro jefe. De igual manera, la gran cacería puede convertirse en la escritura de un libro sobre, pongamos por caso, la mente; y el intento por reproducirse puede convertirse en la búsqueda del reconocimiento público o en la decoración de nuestro apartamento. (Bueno, esta última motivación no se encuentra siempre oculta).

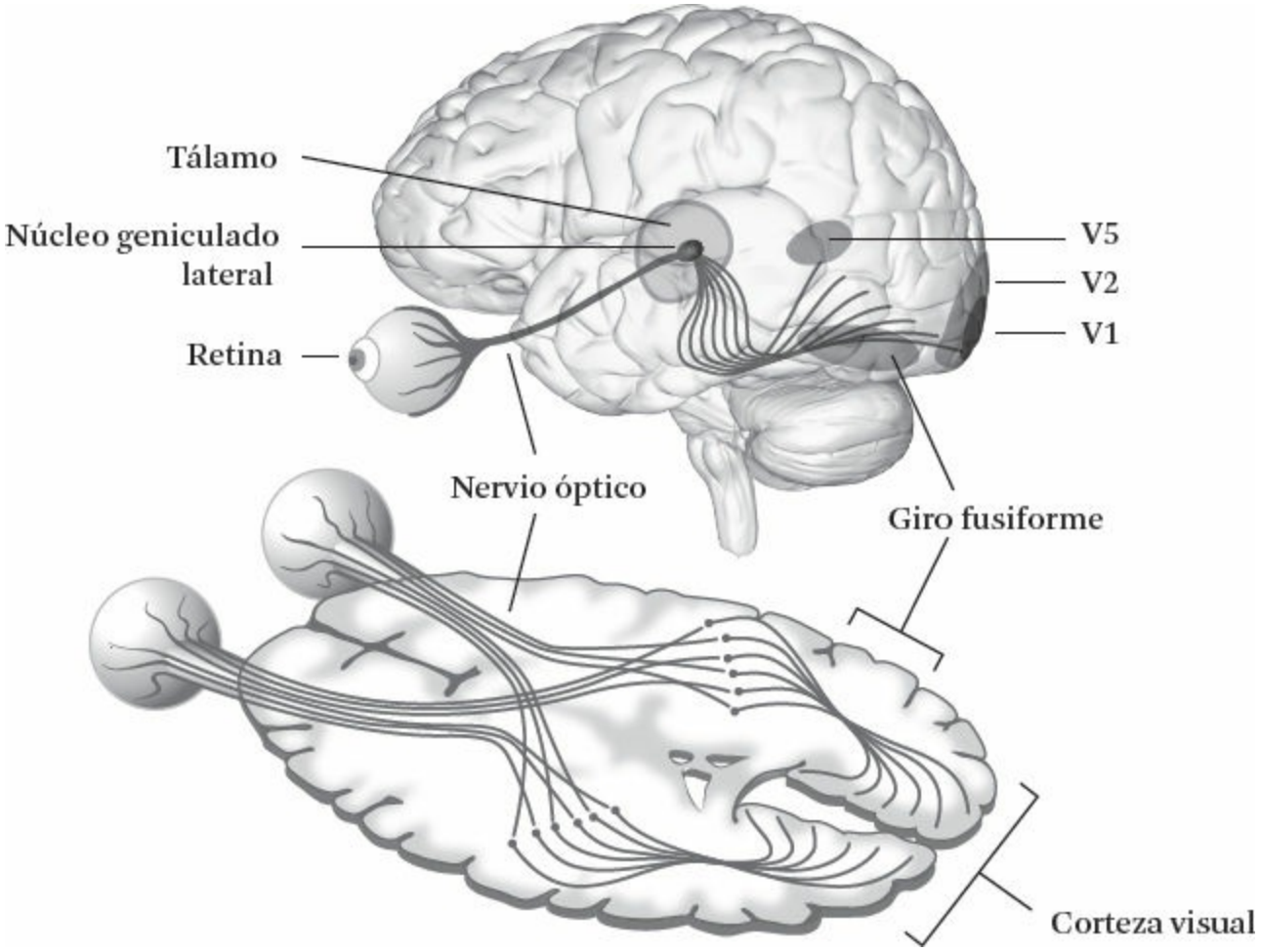
Asimismo, el neocórtex es muy bueno a la hora de ayudarnos a resolver problemas, ya que puede modelizar el mundo de forma precisa reflejando su verdadera naturaleza jerárquica. Sin embargo, es el cerebro antiguo el que nos muestra dichos problemas. Por otra parte, es evidente que, como en cualquier administración eficiente, la manera en que el neocórtex se enfrenta a menudo a los problemas que le surgen se basa en la redefinición de dichos problemas. A propósito de esto último, revisemos el procesamiento de información en el cerebro antiguo.

## *La vía sensitiva*

> Las imágenes, propagadas por el movimiento a lo largo de las fibras de los nervios ópticos del cerebro, son las causantes de la visón.
>
> —ISAAC NEWTON

> Todos nosotros vivimos dentro del universo (o prisión) de nuestro propio cerebro. A partir de él se proyectan millones de frágiles fibras nerviosas sensitivas reunidas en grupos adaptados de forma única. Estos grupos sirven para tomar muestras de los estados energéticos del mundo que nos rodea, es decir, de su calor, de su luz, de su fuerza y de su composición química. Eso es todo lo que podemos llegar a saber directamente del mundo, todo lo demás es inferencia lógica.
>
> —VERNON MOUNTCASTLE (1)

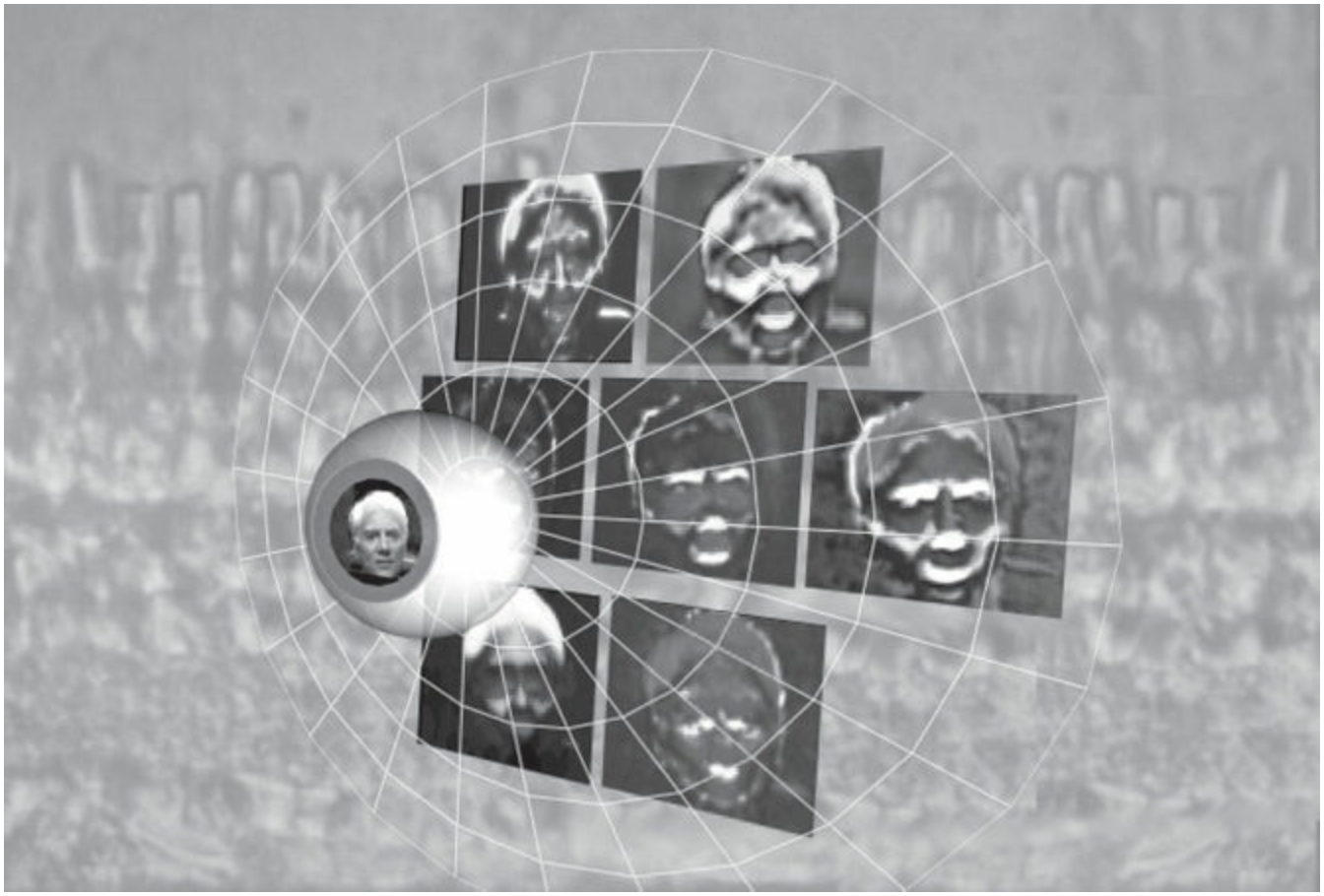
Aunque experimentamos la ilusión de recibir imágenes de alta resolución desde los ojos, lo que el nervio óptico realmente envía al cerebro es tan solo una serie de bocetos y de indicios sobre los puntos de interés en nuestro campo visual. Después lo que hacemos básicamente es tener una alucinación del mundo a partir de recuerdos corticales que interpretan una serie de películas dotadas de una tasa de transferencia de datos muy baja que es recibida a través de canales en paralelo. En una investigación publicada en *Nature*, Frank S. Werblin, profesor de biología molecular y celular en la *University of California at Berkeley*, y el médico y estudiante de doctorado Boton Roska demostraron que el nervio óptico contiene entre diez y doce canales de output, cada uno de los cuales contiene solamente una pequeña cantidad de información sobre una escena determinada (2). Un grupo de las llamadas células ganglionares envía información únicamente sobre contornos (cambios de contraste). Otro grupo detecta únicamente amplias áreas de color uniforme, mientras que un tercer grupo es sensible solamente a los fondos detrás de las figuras de interés.




**La vía visual del cerebro.**

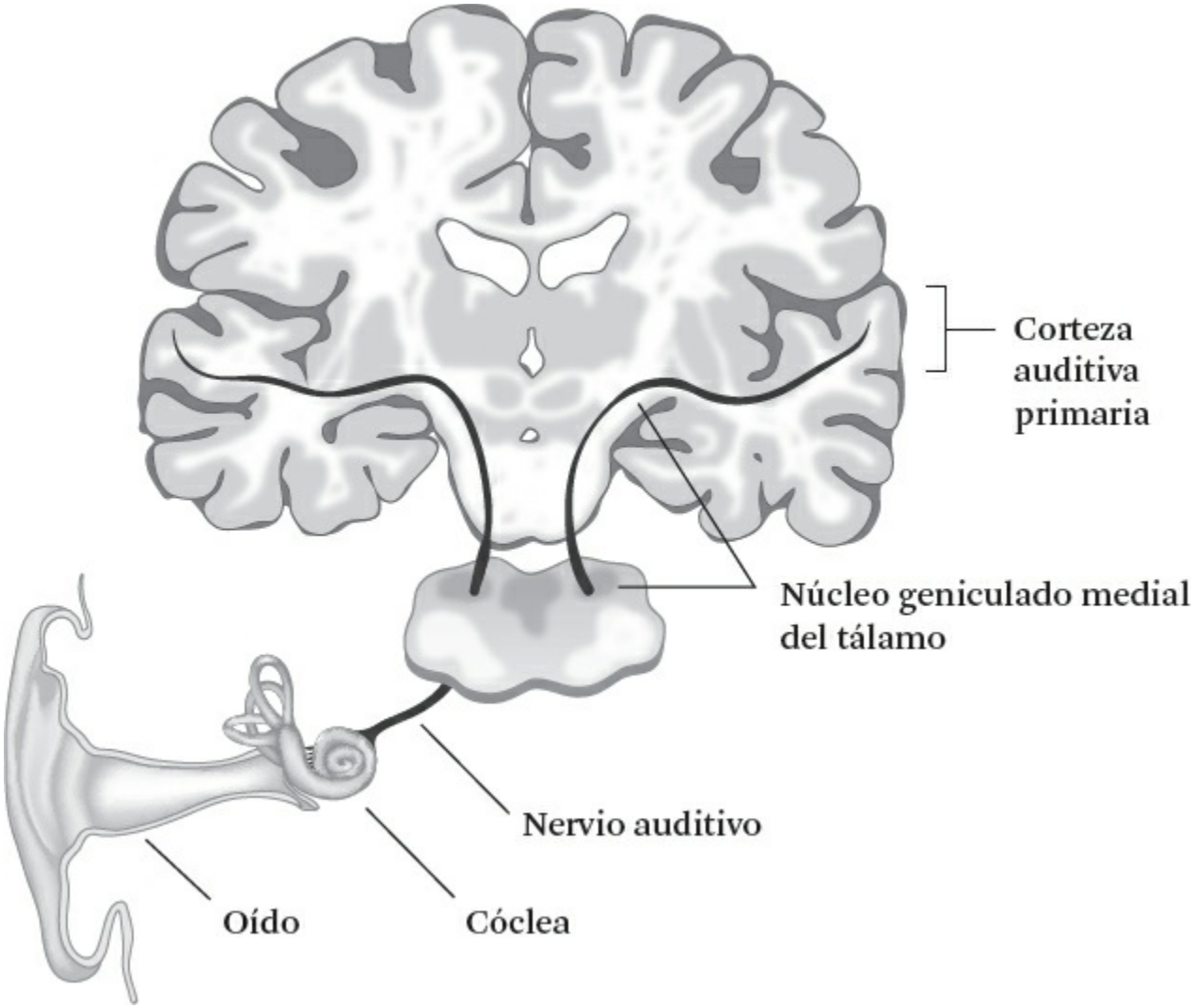
"Aunque creemos ver el mundo de forma plena, lo que recibimos son solo retazos, contornos en el espacio y en el tiempo", dice Werblin. "Estas 12 imágenes del mundo constituyen todas la información que nos es posible obtener sobre lo que hay ahí fuera, y a partir de estas 12 imágenes tan dispersas reconstruimos la riqueza del mundo visual. Me pregunto cómo seleccionó la naturaleza estas 12 sencillas películas y cómo es posible que sean suficientes para proporcionarnos toda la información que parecemos necesitar".

Esta reducción de datos es lo que en el campo de la IA llamamos "codificación dispersa"*. Al crear sistemas artificiales hemos descubierto que desechar la mayor parte del input informativo y quedarse solo con los detalles más importantes proporciona mejores resultados. Si no la limitada capacidad para procesar información que tiene un neocórtex (ya sea este biológico o no) se ve desbordada.



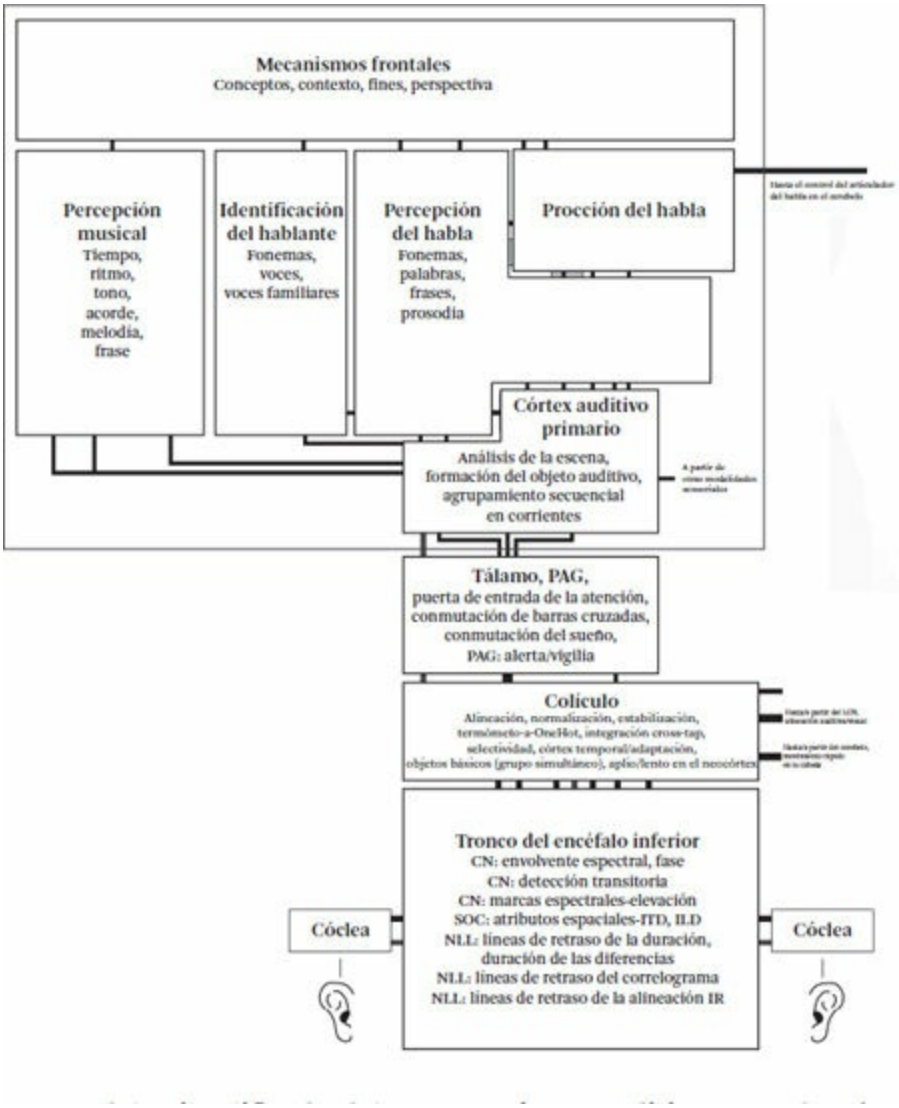
**Siete de las doce "películas" de baja transmisión de datos enviadas por el nervio óptico hasta el cerebro.**

El procesamiento de la información auditiva procedente de la cóclea que atraviesa las regiones subcorticales y luego también atraviesa las primeras capas del neocórtex, ha sido meticulosamente modelizada por Lloyd Watts y su equipo de investigación en *Audience, Inc*. (3). Allí han desarrollado tecnologías para la investigación capaces de extraer 600 bandas de frecuencia diferentes (60 por octava) a partir del sonido. Esto se acerca mucho a la estimación que sostiene que la cóclea humana es capaz de extraer 3 000 bandas (por su parte, los sistemas comerciales de reconocimiento del habla solo utilizan entre 16 y 32 bandas). Mediante dos micrófonos y su detallado modelo de alta resolución espectral del procesamiento auditivo, *Audience* ha creado una tecnología comercial dotada de una resolución espectral un tanto menor que la de su sistema de investigación. Esta tecnología comercial elimina de forma efectiva el ruido de fondo en las conversaciones. Ya se está utilizando esto en muchos teléfonos móviles muy populares y se trata de un asombroso ejemplo de un producto comercial basado en la comprensión de cómo el sistema auditivo humano es capaz de centrarse en una fuente de sonido considerada de interés.




**La vía auditiva del cerebro.**

Inputs procedentes del cuerpo (que se estima que se reciben a cientos de megabits por segundo e incluyen los inputs nerviosos procedentes de la piel, músculos, órganos y otras áreas) son recibidos en la parte superior de la médula espinal. Estos mensajes incluyen más que meras comunicaciones sobre el tacto, también portan información sobre la temperatura, los niveles de ácido (por ejemplo, del ácido láctico de los músculos), el movimiento de la comida a través del tracto gastrointestinal y sobre muchas otras señales. Estos datos son procesados por el tronco del encéfalo y por el mesencéfalo. Unas células de vital importancia llamadas neuronas de la lámina 1 crean un mapa del cuerpo que representa su estado actual. Su función no es muy diferente a la de los *displays* utilizados por los controladores aéreos para seguir a los aviones. A partir de aquí, los datos sensoriales se dirigen a una región enigmática llamada tálamo que nos va a llevar hasta nuestro siguiente tema.



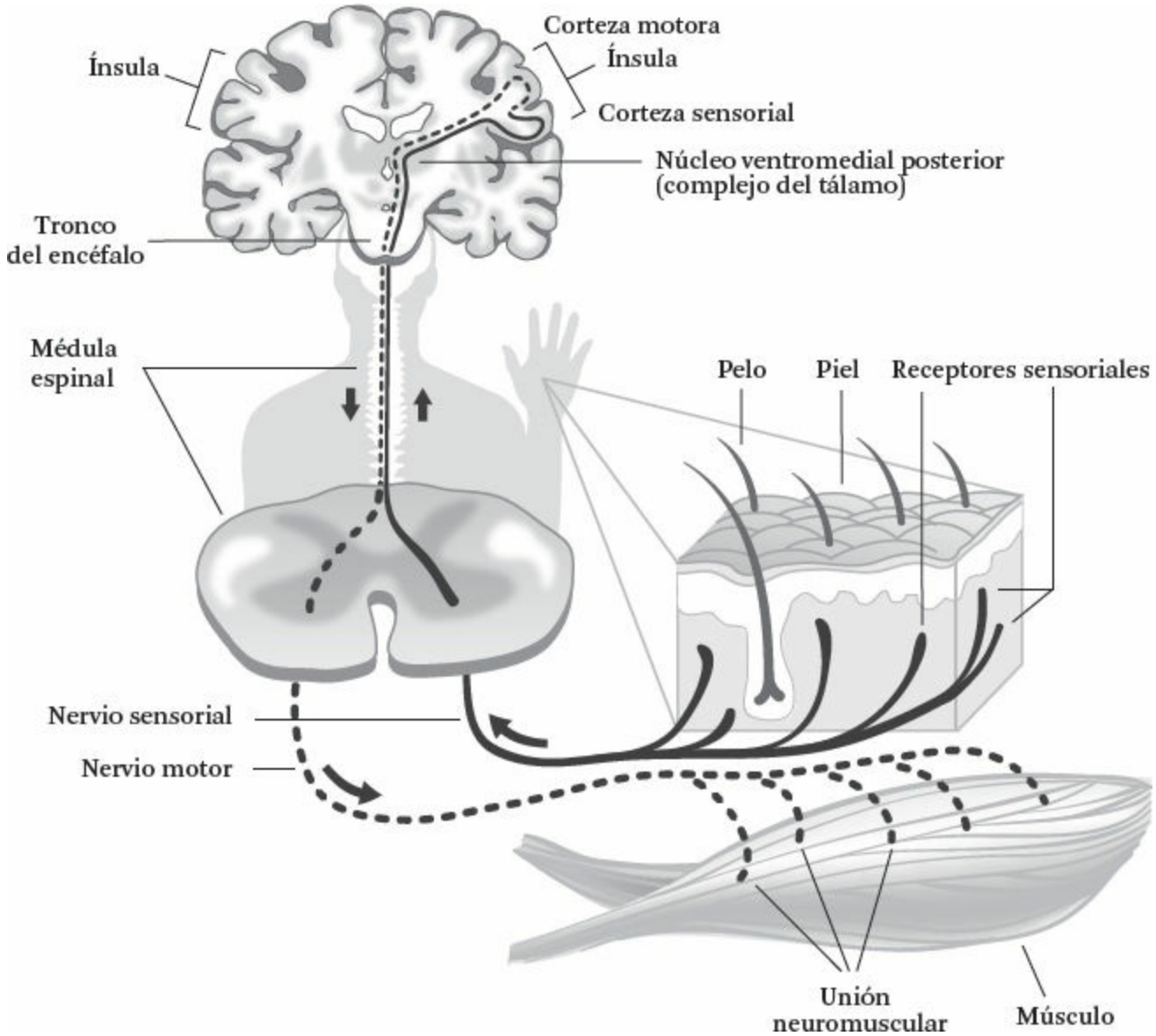

**Un modelo simplificado del procesamiento auditivo en ambas áreas subcorticales (las áreas anteriores al neocórtex) y el neocórtex. Creado por *Audience, Inc*. Imagen adaptada a partir de "*Reverse-Engineering the Human Auditory Pathway*", de L. Watts; en J. Liu et al. (eds.), *WCCI 2012* (Berlin: Springer-Verlag, 2012), p. 49.**

# *El tálamo*

> Todo el mundo sabe lo que es la atención. Es la toma de posesión, clara y distinta, por parte de la mente de uno de los aparentemente varios objetos o trenes de pensamiento simultáneamente posibles. Su esencia son la focalización y la concentración de la consciencia. Conlleva la retirada de ciertas cosas para poder encargarse de otra de forma efectiva.
>
> —WILLIAM JAMES

Desde el mesencéfalo, la información sensorial fluye a través de una región del tamaño de una nuez llamada núcleo ventromedial posterior (VMpo) del tálamo. Esta región calcula reacciones complejas en términos de estados corporales tales como "esto sabe fatal", "qué pestilencia" o "esa iluminación es estimulante". El aumento de la cantidad de información procesada desemboca en dos regiones del neocórtex

llamadas ínsulas. Estas estructuras del tamaño de dedos meñiques se encuentran a la izquierda y a la derecha del neocórtex. El Dr. Arthur Craig, del *Barrow Neurological Institute* de Phoenix, describe el VMpo y las dos regiones insulares como "un sistema que representa el yo material" (4).




**La vía tacto-sensorial del cerebro.**

Además de todas sus otras funciones, el tálamo se considera como una puerta de acceso para que la información sensorial preprocesada pueda acceder al neocórtex. Además de la información táctil que fluye por el VMpo, la información procesada procedente del nervio óptico (que, tal y como se ha dicho anteriormente, ya se ha visto sustancialmente transformada) es enviada hasta una región del tálamo llamada núcleo geniculado lateral, que es quien la envía hasta la región V1 del neocórtex. La información procedente del sentido auditivo pasa a través del núcleo geniculado medial del tálamo en dirección a las primeras regiones auditivas del neocórtex. La totalidad de nuestros datos sensoriales, con la excepción del sistema olfativo que aparentemente utiliza el bulbo olfativo en su lugar, atraviesa regiones específicas del

tálamo.

Sin embargo, el papel más importante que desempeña el tálamo es la comunicación permanente con el neocórtex. Los reconocedores de patrones en el neocórtex envían los resultado provisionales hasta el tálamo y reciben respuestas principalmente mediante las señales recíprocas tanto excitativas como inhibidoras procedentes de la capa VI de cada reconocedor. Téngase en cuenta que no se trata de mensajes inalámbricos, de manera que es necesario contar con una gran cantidad de cableado en forma de axones que recorra todas las regiones del neocórtex y del tálamo. Consideremos la enorme cantidad de mobiliario (medido en términos de masa física requerida por las conexiones) necesario para que los cientos de millones de reconocedores de patrones del neocórtex puedan ser registrados en el tálamo (5) de forma permanente.

¿Qué es lo que le cuentan al tálamo los cientos de millones de reconocedores de patrones neocorticales? Parece tratarse de una conversación importante, ya que un daño importante en la región principal del tálamo puede provocar un prolongado estado bilateral de inconsciencia. Una persona con el tálamo dañado puede que todavía tenga actividad en el neocórtex, en el sentido de que es posible que los pensamientos autogenerados mediante asociación sigan funcionando. Sin embargo, el pensamiento directo, el tipo de pensamiento que nos saca de la cama, nos mete en el coche y nos hace sentarnos frente a la mesa de trabajo, no puede funcionar sin el tálamo. Un caso famoso es el de Karen Ann Quinlan, quien con veintiún años sufrió un ataque al corazón y una parada respiratoria, y que se mantuvo en un estado aparentemente vegetativo e inconsciente durante diez años. Al morir, su autopsia reveló que su neocórtex era normal pero que su tálamo había sido destruido.

Para desempeñar este papel fundamental en nuestra capacidad de centrar la atención, el tálamo depende del conocimiento estructurado contenido en el neocórtex. Así, es capaz de seguir paso a paso una lista que esté almacenada en el neocórtex para permitirnos seguir un tren de pensamiento o un plan de acción. Aparentemente, somos capaces de mantener al mismo tiempo hasta cuatro asuntos en nuestra memoria de trabajo. Esto corresponde a dos por hemisferio, según una reciente investigación llevada a cabo por neurocientíficos del *Picower Institute for Learning and Memory* del MIT (6). La cuestión de si el tálamo controla al neocórtex o viceversa está lejos de verse aclarada. Sin embargo, lo cierto es que sin ambos no podemos funcionar.

## *El hipocampo*

Cada hemisferio del cerebro contiene un hipocampo, una pequeña región que parece

un caballito de mar incrustado en el lóbulo temporal medio. Su función principal es recordar hechos recientes. Dado que la información sensorial fluye a través del neocórtex, depende del neocórtex el determinar si una experiencia es reciente. Si así fuera, tendría que proceder a presentarla ante el hipocampo. Se comporta así bien por su incapacidad para reconocer un conjunto determinado de características (por ejemplo, una cara nueva) o porque se da cuenta de que una situación por lo demás habitual posee ciertos atributos distintivos (como por ejemplo ocurriría si su cónyuge se pusiera un bigote postizo).

El hipocampo es capaz de recordar estas situaciones, aunque parece hacerlo sobre todo mediante indicaciones en el interior del neocórtex. Así, los recuerdos del hipocampo también se almacenan a modo de patrones de nivel más bajo que ya fueron reconocidos y almacenados anteriormente en el neocórtex. Para que los animales sin neocórtex puedan modular las experiencias sensoriales, el hipocampo se limitará a recordar la información procedente de los sentidos, aunque esta habrá tenido que experimentar un preprocesamiento sensorial (por ejemplo, las transformaciones realizadas por el nervio óptico).

Aunque el hipocampo hace uso del neocórtex (si este último está presente) como si fuera su bloc de apuntes, su memoria (compuesta de indicios en el interior del neocórtex) no es intrínsecamente jerárquica.

En consecuencia, los animales que no tienen neocórtex pueden recordar cosas usando su hipocampo, pero sus recuerdos no son jerárquicos. Además, la capacidad del hipocampo es limitada, de manera que su memoria es a corto plazo. Transfiere una secuencia concreta de patrones desde su memoria a corto plazo hasta la memoria jerárquica a largo plazo del neocórtex después de reproducir esta secuencia de recuerdos ante el neocórtex una y otra vez. Por tanto, necesitamos un hipocampo para aprender nuevos recuerdos y capacidades, aunque las capacidades estrictamente motoras parecen usar un mecanismo diferente. Alguien con ambas copias de su hipocampo dañadas es capaz de retener sus recuerdos ya existentes, pero no es capaz de crear recuerdos nuevos.

El neurocientífico Theodore Berger y sus colegas de la *University of Southern California* han modelizado el hipocampo de una rata y han tenido éxito a la hora de implantarles a las ratas un hipocampo artificial. Mediante drogas, en una investigación del año 2011 los científicos de la USC bloquearon en las ratas ciertos comportamientos aprendidos. Utilizando un hipocampo artificial, las ratas fueron capaces de volver a aprender el comportamiento rápidamente. “Se activa el interruptor y las ratas se acuerdan. Se desactiva y las ratas se olvidan”, escribió Berger haciendo referencia a su capacidad de controlar a distancia los implantes neuronales. En otro experimento, el científico permitió que sus hipocampos artificiales funcionasen junto a los hipocampos naturales de las ratas. El resultado fue que la capacidad de las ratas para aprender nuevos comportamientos se reforzó.

“Estas investigaciones de modelización experimental integrada demuestran por primera vez”, explicó Berger, “que […] una prótesis neuronal capaz de manipular e identificar el proceso de codificación puede restaurar e incluso mejorar los procesos nemotécnicos cognitivos” (7). El hipocampo es una de las regiones que primero se deterioran a causa del Alzheimer, de manera uno de los objetivos de esta investigación es desarrollar un implante neuronal para humanos que mitigue la primera fase del daño provocado por la enfermedad.

## *El cerebelo*

Existen dos estrategias que nos permiten capturar una pelota al vuelo. Se pueden resolver las ecuaciones diferenciales simultáneas y complejas que controlan el movimiento de la pelota, así como las otras ecuaciones que controlan nuestro propio ángulo de visión de la pelota, y luego calcular todavía más ecuaciones para averiguar cómo mover el cuerpo, el brazo y la mano para que estén en el sitio adecuado en el momento adecuado.

Esta no es la estrategia por la que opta nuestro cerebro. Básicamente, él lo que hace es simplificar el problema colapsando muchas ecuaciones en un simple modelo tendencial que tenga en cuenta las tendencias sobre dónde parece estar la pelota en nuestro campo de visión y cuál es la rapidez con la que la pelota se está moviendo dentro de él. El cerebro hace lo mismo para la mano. Realiza predicciones lineales de la posición aparente de la pelota en nuestro campo de visión y de la posición aparente de la mano. Por supuesto, el objetivo es asegurarse de que se encuentren en el mismo punto del espacio y del tiempo. Si la pelota parece estar cayendo demasiado rápido y la mano parece estar moviéndose demasiado despacio, el cerebro dará la orden a la mano para que se mueva más rápido, de manera que las tendencias coincidan. Esta solución de “nudo gordiano” a lo que de otra manera sería un problema matemático inabarcable recibe el nombre de funciones base. Estas funciones son llevadas a cabo por el cerebelo, una región en forma de judía y de un tamaño aproximado al de una pelota de béisbol, que está situada en el tronco del encéfalo (8).

El cerebelo es una región del cerebro antiguo que en el pasado controló prácticamente todos los movimientos de los homínidos. Todavía contiene la mitad de las neuronas del cerebro, aunque la mayoría son relativamente pequeñas, de manera que la región constituye solamente alrededor del 10 % del peso del cerebro. Asimismo, el cerebelo es otro ejemplo de la enorme cantidad de repeticiones que encontramos en el diseño del cerebro. Sobre su diseño en el genoma existe relativamente poca información, ya que su estructura es un patrón de varias neuronas

que se repite miles de millones de veces. Además, al igual que ocurre con el neocórtex, existe una uniformidad a lo largo y ancho de su estructura (9).

La mayor parte de la función que controla nuestros músculos ha sido asumida por el neocórtex. Para ello ha utilizado los mismos algoritmos de reconocimiento de patrones que usa para la percepción y cognición. En el caso del movimiento, la mejor manera de caracterizar la función del neocórtex es como la de implementador de patrones. El neocórtex hace uso de la memoria en el cerebelo para registrar delicadas secuencias de movimientos, por ejemplo la firma y ciertos ademanes ostentosos de la expresión artística en la música y el baile. Estudios sobre el papel del cerebelo durante el aprendizaje de la escritura en los niños revelan que las células de Purkinje en el cerebelo toman muestras de las secuencias de movimientos y que cada una de estas células es sensible a una muestra en concreto (10). Dado que ahora la mayor parte de nuestros movimientos están controlados por el neocórtex, mucha gente puede arreglárselas con una discapacidad en el cerebelo que sea relativamente pequeña o incluso con un daño en el mismo que sea significativo. Lo único que pasa entonces es que sus movimientos pueden volverse menos gráciles.

El neocórtex también puede recurrir al cerebelo con objeto de utilizar su capacidad para calcular funciones base en tiempo real que anticipen los resultados de las acciones que estamos sopesando pero que todavía no hemos realizado (y puede que nunca realicemos), así como las acciones o posibles acciones de otros. Se trata de otro ejemplo de los innatos predictores lineales integrados en el cerebro.

Se ha logrado un gran progreso a la hora de simular el cerebelo en lo que respecta a su capacidad para responder dinámicamente ante indicios sensoriales. Para ello se utilizan las funciones base que he comentado anteriormente en simulaciones de abajo a arriba. Como base se utilizan modelos bioquímicos y simulaciones de arriba a abajo construidas según modelos matemáticos del funcionamiento de cada unidad que se repite en el cerebelo (11).

## *Placer y miedo*

> El miedo es la fuente principal de superstición y una de las principales fuentes de crueldad. La superación del miedo es el origen de la sabiduría.
>
> —BERTRAND RUSSELL

> Siente miedo, pero hazlo de todas maneras.
>
> —SUSAN JEFFERS

Si el neocórtex hace un buen trabajo a la hora de resolver problemas, ¿cuál es el problema principal que estamos intentando resolver? El problema que desde siempre

ha intentado resolver la evolución es la supervivencia de las especies. Eso se traduce en la supervivencia del individuo. Así, cada uno de nosotros utiliza su propio neocórtex para interpretar ese hecho en un sinnúmero de formas posibles. Para sobrevivir, los animales necesitan obtener su siguiente comida a la vez que evitan convertirse en la comida de otro. También tienen que reproducirse. Los primeros cerebros dieron lugar a los sistemas del placer y del miedo que recompensaban el cumplimiento de estas necesidades básicas a la vez que daban lugar a las conductas elementales que las hacen posibles. A medida que los medios y las especies en liza cambiaban gradualmente, la evolución biológica realizaba sus correspondientes cambios. Con la aparición del pensamiento jerárquico, el cumplimiento de los impulsos fundamentales se hizo más complejo, ya que ahora estaba sujeto al enorme conjunto de ideas contenido en las propias ideas. Sin embargo, pese al importante ajuste llevado a cabo por el neocórtex, el cerebro antiguo sigue vivito y coleando y nos sigue motivando mediante el placer y el miedo.

Una región que se asocia con el placer es el núcleo accumbens. En un famoso experimento realizado en la década de 1950, ratas capaces de estimular directamente esta pequeña región presionando una palanca que activaba electrodos previamente implantados, preferían llevar a cabo esta estimulación antes que cualquier otra cosa, incluyendo las relaciones sexuales o el comer, y acababan por agotarse y dejarse morir de inanición (12). En humanos, otras regiones también tienen que ver con el placer, como por ejemplo el pálido ventral y, por supuesto, el propio neocórtex.

El placer también viene regulado por productos químicos como la dopamina y la serotonina. La exposición en detalle de este sistema es algo que está más allá del alcance de este libro, sin embargo es importante tomar consciencia de que hemos heredado estos mecanismos de nuestros primos los antecesores de los mamíferos. La función del neocórtex es permitirnos ser los dueños del placer y del miedo, no sus esclavos. En la medida en que a menudo nos vemos sometidos a comportamientos adictivos, el neocórtex no siempre tiene éxito en su empeño. Concretamente, la dopamina es un neurotransmisor involucrado en la experimentación del placer. Si nos pasa algo bueno (ganamos en la lotería, conseguimos el reconocimiento de nuestros colegas, nos da un abrazo un ser querido o simplemente conseguimos un pequeño logro como el de hacer que un amigo se ría de un chiste) experimentamos una segregación de dopamina. Al igual que las ratas que murieron sobreestimulando su núcleo accumbens, a veces optamos por un atajo para lograr estos estallidos de placer, lo cual no es siempre una buena idea.

Los juegos de azar, por ejemplo, pueden hacer que segreguemos dopamina (por lo menos cuando ganamos) pero esto está sujeto a su inherente falta de predictibilidad. Los juegos de azar pueden dar resultado a la hora de segregar dopamina durante un tiempo, pero dado que las posibilidades de ganar están intencionadamente organizadas en nuestra contra (de lo contrario el modelo de

negocio de los casinos no funcionaría) el juego puede convertirse en algo ruinoso si se adopta como estrategia habitual. Peligros similares vienen asociados a cualquier comportamiento adictivo. Una mutación genética determinada del gen D2 receptor de dopamina causa sentimientos de placer especialmente fuertes a partir de la experimentación inicial con sustancias y comportamientos adictivos, pero también se sabe con certeza (aunque no siempre se tome nota de ello) que la capacidad de estas sustancias para producir placer disminuye gradualmente con su uso. Otra mutación genética produce que las personas no segreguen niveles normales de dopamina a partir de los logros cotidianos, lo cual también puede conllevar que se busque el aumento del placer en las primeras experiencias con las actividades adictivas. La minoría de la población que sufre estas propensiones genéticas hacia la adicción da lugar a un enorme problema social y médico. Incluso aquellos que logran evitar graves comportamientos adictivos tienen que debatirse entre equilibrar las recompensas de la segregación de dopamina y las consecuencias de los comportamientos que las segregan.

La serotonina es un neurotransmisor que juega un importante papel en la regulación del estado de ánimo. A niveles altos se asocia a sentimientos de bienestar y felicidad. Además, la serotonina tiene otras funciones, incluyendo la potencia sináptica, el apetito, el sueño, el deseo sexual y la digestión. Medicamentos antidepresivos como los inhibidores selectivos de la recaptación de serotonina, que tienden a aumentar los niveles de serotonina al alcance de los receptores, son propensos a producir efectos de gran alcance, no todos deseables (como por ejemplo la supresión de la libido). A diferencia de las acciones que tienen lugar en el neocórtex, donde el reconocimiento de patrones y la activación de axones afectan solamente a un pequeño número de circuitos neocorticales cada vez, estas sustancias tienen efectos sobre extensas regiones del cerebro o incluso sobre el sistema nervioso en su conjunto.

Cada hemisferio del cerebro humano posee una amígdala, que consiste en una región en forma de almendra que contiene varios lóbulos pequeños. La amígdala también forma parte del cerebro antiguo y está involucrada en el procesamiento de ciertos tipos de respuestas emocionales, la más importante de las cuales es el miedo. En los antecesores de los mamíferos, ciertos estímulos preprogramados que simbolizan al peligro son transmitidos directamente hasta la amígdala, que a su vez desencadena el mecanismo de “lucha o huída”. En los humanos, la amígdala depende de que las percepciones de peligro sean transmitidas por el neocórtex. Por ejemplo, un comentario negativo hecho por nuestro jefe puede desencadenar una respuesta así generando miedo a perder nuestro trabajo (o puede que no, depende de si confiamos en un plan B). Una vez que la amígdala decide que nos enfrentamos a un peligro, tiene lugar una secuencia de sucesos ancestral. La amígdala da la señal para que la

glándula pituitaria segregue una hormona llamada ACTH (adrenocorticotropa). A su vez, esto desencadena la hormona del estrés llamada cortisol procedente de las glándulas suprarrenales, lo cual produce que nuestros músculos y sistema nervioso reciban una mayor cantidad de energía. Las glándulas suprarrenales también producen adrenalina y noradrenalina, que reprimen nuestros sistemas digestivos, inmunológicos y reproductivos, ya que consideran que en un caso de emergencia no son procesos prioritarios. Los niveles de presión sanguínea, azúcar en sangre, colesterol y fibrinógeno (que acelera la coagulación de la sangre) aumentan. El ritmo cardiaco y la respiración se aceleran. Incluso las pupilas llegan a dilatarse para tener una mayor agudeza visual sobre el enemigo o sobre la vía de escape. Esto es muy beneficioso ante un peligro real como puede serlo que de repente un depredador se nos cruce en el camino. Sin embargo, es bien sabido que en el mundo de hoy en día la activación crónica de este mecanismo de lucha o huida puede producir daños permanentes de salud, como por ejemplo hipertensión, altos niveles de colesterol y otros problemas.

El sistema de los niveles generales de los neurotransmisores, como el de la serotonina, y de los niveles hormonales, como el de la dopamina, es un tema complejo al que podríamos dedicar el resto de este libro (siguiendo lo hecho por una gran cantidad de escritos), pero sí que es necesario señalar que el ancho de banda de la información (el ritmo del procesamiento de la información) en este sistema es muy bajo comparado con el ancho de banda del neocórtex. Solo existe un número limitado de sustancias involucradas y los niveles de estos productos químicos tienden a variar lentamente. Además, son relativamente universales por todo el cerebro. Esto se contrapone a lo que pasa en el neocórtex, que está compuesto de cientos de miles de millones de conexiones que pueden cambiar rápidamente.

Es justo decir que nuestras experiencias emocionales tienen lugar tanto en el cerebro antiguo como en el moderno. El pensamiento tiene lugar en el moderno (el neocórtex), pero los sentimientos tienen lugar en ambos. Por tanto, cualquier emulación del comportamiento humano necesitaría modelizar los dos. Sin embargo, si lo que se persigue es solo la inteligencia cognitiva humana, el neocórtex es suficiente. Podemos remplazar el cerebro antiguo con la motivación más directa producida por un neocórtex no biológico para así alcanzar los objetivos que le asignemos. Por ejemplo, en el caso de Watson el objetivo era muy simple: responder correctamente a las preguntas de *Jeopardy!* (y no obstante que estas respuestas se ajustaran a un programa que comprendiera el sistema de apuestas de *Jeopardy!*). En el caso del nuevo sistema de Watson, que está siendo desarrollado conjuntamente por Nuance e IBM y que está orientado hacia el conocimiento médico, el objetivo es ayudar a tratar enfermedades humanas. Los sistemas futuros podrán tener objetivos como la cura de las enfermedades y la lucha contra la pobreza. En el caso de los humanos, gran parte de la batalla entre placer y miedo ya se ha quedado obsoleta, ya

que el cerebro antiguo evolucionó mucho antes de que las sociedades humanas primitivas ni siquiera fueran fundadas. De hecho, proviene en gran parte de los reptiles.

Existe una lucha permanente en el cerebro humano sobre quién es el que manda, el cerebro moderno o el antiguo. El cerebro antiguo trata de imponer una agenda basada en su control sobre las experiencias de placer y de miedo, mientras que el cerebro moderno intenta permanentemente comprender los algoritmos relativamente primitivos del cerebro antiguo e intenta manipularlo en favor de su propia agenda. Téngase en cuenta que la amígdala es incapaz de evaluar el peligro por sí sola, ya que el cerebro humano depende del neocórtex para llevar a cabo dichos juicios. ¿Es esa persona un amigo o un enemigo, alguien querido o una amenaza? Eso solo lo puede decidir el neocórtex.

En la medida en la que no nos embarquemos directamente en un combate mortal o en tener que cazar para poder comer, habremos tenido éxito a la hora de sublimar, por lo menos parcialmente, nuestras primitivas tendencias en comportamientos que conllevan una mayor creatividad. Partiendo de esta base, analizaremos en el siguiente capítulo las cuestiones de la creatividad y del amor.

* *Sparse coding* en el original.

CAPÍTULO SEIS

# Capacidades transcendentes

He aquí mi sencilla religión. No hay necesidad de templos ni de una complicada filosofía. Nuestro propio cerebro, nuestro propio corazón, he ahí nuestro templo; la filosofía de la bondad.

—EL DALAI LAMA

Mi mano se mueve porque ciertas fuerzas, la eléctrica, la magnética o cualquier "fuerza nerviosa" que pudiera existir, están impresas en mi cerebro. Es probable que, si la Ciencia fuera algo acabado, se pudiera rastrear esta fuerza nerviosa almacenada en mi cerebro hasta llegar a las fuerzas químicas que llegan al cerebro por medio de la sangre y que en último término se derivan de la comida que como y del aire que respiro.

—LEWIS CARROLL

Nuestros pensamientos emocionales también tienen lugar en el neocórtex, pero están bajo la influencia de porciones del cerebro que van desde regiones cerebrales primitivas como las amígdalas hasta algunas estructuras cerebrales evolutivamente recientes como las neuronas en huso, que parecen jugar un papel fundamental en las emociones de nivel más alto. A diferencia de las recursivas estructuras lógicas y regulares encontradas en el córtex cerebral, las neuronas en huso tienen formas y conexiones muy irregulares. Son las neuronas más grandes del cerebro humano y abarcan toda su extensión; y además están profundamente interconectadas mediante cientos de miles de conexiones que mantienen unidas diferentes partes del neocórtex.

Tal y como mencioné anteriormente, la ínsula ayuda a procesar las señales sensoriales, sin embargo también desempeña un papel fundamental en las emociones de más alto nivel. Las células fusiformes se originan a partir de esta región. Los escáneres de imagen por resonancia magnética funcional (fMRI) han revelado que estas células son especialmente activas cuando una persona se enfrenta a emociones tales como el amor, el enfado, la tristeza y el deseo sexual. Las situaciones que las activan especialmente también incluyen los momentos en los que un sujeto mira a su pareja u oye llorar a su hijo o hija.

Las células fusiformes tienen largos filamentos neuronales llamados dendritas apicales, que son capaces de conectarse a regiones neocorticales alejadas. Esta "profunda" interrelación en la cual ciertas neuronas proporcionan conexiones a lo

largo de diversas regiones es una característica que a medida que ascendemos por la escala evolutiva tiene lugar con más frecuencia. No es de sorprender que las células en huso, al estar involucradas en el manejo de la emoción y del juicio moral, posean este tipo de interrelación, dada la capacidad de las reacciones emocionales de más alto nivel de tocar diversos temas y pensamientos. Debido a sus conexiones con muchas otras partes del cerebro, las emociones de nivel alto que procesan las células en huso se ven afectadas por todas las regiones perceptivas y cognitivas. Es importante señalar que estas células no resuelven problemas racionales, razón por la cual no tenemos control racional sobre nuestras respuestas ante la música o el enamoramiento. Sin embargo, el resto del cerebro está muy involucrado en intentar dar sentido a nuestras misteriosas emociones de alto nivel.

Existen relativamente pocas células en huso, solo unas 80 000, aproximadamente 45 000 en el hemisferio derecho y 35 000 en el izquierdo. Esta disparidad constituye al menos una razón para explicar la percepción de que la inteligencia emocional es cosa del cerebro derecho, aunque la desproporción existente es más bien modesta. Los gorilas tienen alrededor de 16 000 de estas células, los bonobos unas 2 100 y los chimpancés unas 1 800. El resto de mamíferos carece de ellas por completo.

Los antropólogos creen que las células en huso hicieron su aparición hace 10 o 15 millones de años en el todavía por descubrir ancestro común que comparten los monos y los homínidos (precursores de los humanos), y que el número de estas células aumentó rápidamente hará unos 100 000 años. Curiosamente, las células en huso no existen en los humanos recién nacidos, sino que empiezan a aparecer alrededor de los cuatro meses después del nacimiento e incrementan considerablemente su número entre el primer y el tercer año de vida. La capacidad de los niños para afrontar cuestiones morales y percibir emociones de más alto nivel como por ejemplo en amor también se desarrolla en ese periodo.

## *Aptitud*

Wolfgang Amadeus Mozart (1756–1791) escribió un minueto cuando tenía cinco años. A la edad de seis tocó para la emperatriz María Teresa en la corte imperial de Viena. Compuso 600 piezas, incluyendo 41 sinfonías, antes de morir a la edad de 35 años, y por lo general se le considera como el compositor de la tradición clásica europea más importante de todos los tiempos. Por tanto, se podría decir que tenía una aptitud para la música.

¿Qué significa eso en el contexto de la teoría de la mente basada en el reconocimiento de patrones? Está claro que parte de lo que consideramos como aptitud es producto de la educación, es decir, de las influencias ejercidas por el

medio y por otras personas. Mozart nació en una familia musical. Su padre, Leopold, fue compositor y *kapellmeister* (literalmente, director musical) en la orquesta de la corte del arzobispo de Salzburgo. El joven Mozart estuvo inmerso en música y su padre empezó a enseñarle a tocar el violín y el *clavier* (un instrumento de teclado) a la edad de tres años.

Sin embargo, las influencias del medio por sí solas no son capaces de explicar por completo la genialidad de Mozart. Claramente también existe un componente natural. ¿Qué forma toma este componente? Tal y como escribí en el capítulo 4, diversas regiones del neocórtex han sido optimizadas mediante la evolución biológica para ciertos tipos de patrones. Aunque el algoritmo básico para el reconocimiento de patrones que poseen los módulos es uniforme a lo largo del neocórtex, debido a que ciertos tipos de patrones tienden a fluir a través de regiones determinadas (por ejemplo, las caras lo hacen a través del giro fusiforme), dichas regiones serán mejores a la hora de procesar los patrones asociados. Sin embargo, existen numerosos parámetros que rigen la manera en que se realiza el algoritmo en cada módulo. Por ejemplo, ¿qué nivel de coincidencia es necesario para que un patrón sea reconocido? ¿Cómo se ve modificado ese umbral si un módulo de nivel más alto envía una señal indicando que su patrón es "esperado"? ¿Cómo se consideran los parámetros del tamaño? Estos y otros factores han sido fijados de forma diferente en regiones diferentes en beneficio de ciertos tipos de patrones en particular. Durante nuestro trabajo con métodos similares en el campo de la inteligencia artificial, observamos el mismo fenómeno y usamos simulaciones de la evolución para optimizar estos parámetros.

Si regiones concretas pueden ser optimizadas para diversos tipos de patrones, entonces significa que en los cerebros individuales también variará la capacidad de aprender, reconocer y crear ciertos tipos de patrones. Por ejemplo, un cerebro puede tener una aptitud innata hacia la música al ser más capaz de reconocer los patrones rítmicos o al ser capaz de comprender mejor los arreglos geométricos de las armonías. El fenómeno del oído absoluto (la capacidad de reconocer y de reproducir un tono sin ninguna referencia externa), que está relacionado con el talento musical, parece tener una base genética, aunque la capacidad necesita ser desarrollada, de manera que es probable que sea una combinación entre naturaleza y medio. La base genética del oído absoluto es posible que resida fuera del neocórtex, en el preprocesamiento de la información auditiva, mientras que el aspecto aprendido es posible que resida en el neocórtex.

Existen otras capacidades que tienen que ver con los niveles de competencia que diferencian a alguien mediocre del genio legendario. Las capacidades neocorticales, por ejemplo la capacidad del neocórtex para dominar las señales de miedo que genera la amígdala cuando se enfrenta a la desaprobación, juegan un papel significativo, al igual que lo hacen atributos como la confianza, las capacidades

organizativas y la habilidad para influir en los demás. Una capacidad muy importante a la que me he referido anteriormente es la valentía de llevar adelante ideas que atentan contra la ortodoxia. Invariablemente, las personas a las que consideramos genios llevaron a cabo sus propios experimentos mentales en formas que en un principio no fueron comprendidas o apreciadas por sus iguales. Aunque Mozart obtuvo el reconocimiento durante su vida, la mayor parte de este reconocimiento le llegó después de muerto. Murió pobre, fue enterrado en una fosa común y solo otros dos músicos acudieron a su funeral.

## *Creatividad*

> La creatividad es una droga sin la que no puedo vivir.
>
> —CECIL B. DEMILLE

> El problema no es nunca cómo conseguir pensamientos nuevos e innovadores para la mente, sino cómo deshacerse de los viejos. Todas las mentes son un edificio lleno de muebles antiguos. Vacíe un rincón de su mente y la creatividad lo llenará de inmediato.
>
> —DEE HOCK

> La humanidad puede ser bastante fría con aquellos cuyos ojos ven el mundo de forma diferente.
>
> —ERIC A. BURNS

> La creatividad puede resolver casi cualquier problema. El acto creativo, la derrota del hábito por medio de la originalidad, lo supera todo.
>
> —GEORGE LOIS

Un aspecto fundamental de la creatividad es el proceso de construcción de buenas metáforas, símbolos que representan otras cosas. El neocórtex es una gran máquina de metáforas que explica por qué somos una especie creativa única. Cada uno de los aproximadamente 300 millones de reconocedores de patrones de nuestro neocórtex reconoce y define un patrón, y además le pone un nombre, lo que en el caso de los módulos de reconocimiento de patrones neocorticales se reduce a un axón emergiendo desde un reconocedor de patrones que se disparará cuando dicho patrón sea encontrado. A su debido tiempo, ese símbolo se convierte en parte de otro patrón. Así, cada uno de esos patrones es esencialmente una metáfora. Los reconocedores pueden dispararse hasta 100 veces por segundo, de manera que potencialmente podemos reconocer hasta 30 mil millones de metáforas por segundo. Por supuesto, no todos los módulos se disparan en cada ciclo, pero es justo decir que de hecho reconocemos millones de metáforas por segundo.

Por supuesto, algunas metáforas tienen más importancia que otras. Darwin se dio

cuenta de que la opinión de Charles Lyell, que defendía que cambios muy graduales que parten de un pequeño chorro de agua pueden esculpir grandes cañones, era una poderosa metáfora para explicar cómo un pequeño chorro de pequeños cambios evolutivos durante miles de generaciones puede esculpir grandes cambios en lo que se refiere a la diferenciación entre especies. Por su parte, los experimentos mentales como el que usó Einstein para ilustrar el verdadero significado del experimento de Michelson-Morley son todos ellos metáforas, en el sentido en que son una "cosa que representa o simboliza otra", para citar una definición de diccionario.

¿Encuentra usted alguna metáfora en el soneto 73 de Shakespeare?

*That time of year thou mayst in me behold*
*When yellow leaves, or none, or few, do hang*
*Upon those boughs which shake against the cold,*
*Bare ruined choirs, where late the sweet birds sang.*
*In me thou seest the twilight of such day*
*As after sunset fadeth in the west,*
*Which by and by black night doth take away,*
*Death's second self that seals up all in rest.*
*In me thou seest the glowing of such fire*
*That on the ashes of his youth doth lie,*
*As the death bed whereon it must expire*
*Consumed with that which it was nourished by.*
*This thou perceiv'st, which makes thy love more strong,*
*To love that well which thou must leave ere long.**

En este soneto el poeta utiliza largas metáforas para describir su avanzada edad. Su edad es como el otoño tardío, "cuando las hojas amarillas, o ninguna o pocas, cuelgan". El tiempo atmosférico es frío y los pájaros ya no se pueden sentar en las ramas a las que llama "desnudos coros arruinados". Su edad es como el crepúsculo durante la "puesta de sol en el poniente, el cual poco a poco la noche se lleva". Él es los restos de un fuego "que descansa sobre las cenizas de su juventud". De hecho, todo lenguaje es en último término metáfora, aunque algunas de sus expresiones son más memorables que otras.

La búsqueda de una metáfora es el proceso de reconocimiento de un patrón pese a las diferencias de detalle y contexto, una actividad que realizamos de forma trivial en cada momento de nuestras vidas. Las conjeturas metafóricas que consideramos importantes tienden a producirse en los intersticios de diferentes disciplinas. Sin embargo, atacar esta fundamental fuerza creativa es la tendencia dominante que nos conduce hacia una especialización de las ciencias cada vez mayor (y en el resto de campos ocurre lo mismo). Tal y como escribió el matemático norteamericano Norbert Wiener (1894–1964) en su seminal libro *Cybernetics*, publicado el año en que nací (1948):

> Existen campos en el trabajo científico, tal y como veremos en el desarrollo de este libro, que han sido explorados desde los diferentes flancos de la matemática pura, de la estadística, de la ingeniería eléctrica

y de la neurofisiología, y en ellos cada idea individual recibe un nombre distinto. Esto significa que importantes trabajos se han hecho por triplicado o cuadruplicado, mientras que otros trabajos igual de importantes han sufrido retrasos por la inexistencia en un determinado campo de resultados que en otro campo pueden haberse convertido ya en clásicos.

Son estas regiones fronterizas las que a menudo le ofrecen las más ricas oportunidades al investigador cualificado. Al mismo tiempo, son las más refractarias a las técnicas de bombardeo masivo y de la división del trabajo.

Una técnica que he utilizado en mi propio trabajo para combatir la creciente especialización es la de reunir los expertos que he juntado para un proyecto (por ejemplo, mi trabajo en el reconocimiento del habla incluyó expertos en el habla, lingüistas y expertos en psicoacústica y en el reconocimiento de patrones, por no mencionar a los informáticos). Después les animo a todos a que enseñen al resto del grupo sus técnicas y terminología particulares. Entonces procedemos a desechar toda esa terminología y creamos la nuestra propia. Así, constantemente nos encontramos con metáforas pertenecientes a un campo que resuelven problemas en otro.

Un ratón que encuentra una vía de escape cuando es descubierto por el gato de la casa, cosa que puede hacer incluso si la situación es un tanto diferente a las que se ha tenido que enfrentar antes, está siendo creativo. Los órdenes de magnitud de nuestra propia creatividad son enormemente superiores a los del ratón y conllevan muchos más niveles de abstracción, ya que tenemos un neocórtex mucho más grande que es capaz de abarcar niveles jerárquicos muy superiores. De manera que una forma de alcanzar una mayor creatividad es ensamblar de forma efectiva mayores cantidades de neocórtex.

Una estrategia para expandir el neocórtex disponible es la colaboración entre múltiples humanos. Esto se consigue de forma rutinaria mediante la comunicación entre personas reunidas en comunidad para resolver un problema. Así, recientemente se han realizado esfuerzos para utilizar herramientas de colaboración online y llevar a cabo colaboraciones en tiempo real. Estas iniciativas han tenido éxito en las matemáticas y en otros campos (1).

Por supuesto, el siguiente paso será expandir el propio neocórtex mediante un equivalente no biológico. Este será nuestro acto creativo definitivo: crear la capacidad de ser creativo. En último término, el neocórtex no biológico se volverá más rápido que el biológico y podrá encontrar más rápidamente el tipo de metáforas que inspiraron a Darwin y Einstein. Así, podría explorar de forma sistemática el solapamiento de todas las fronteras en crecimiento exponencial que separan nuestros diferentes campos de conocimiento.

Algunas personas se muestran preocupadas sobre lo que les pasará a aquellos que opten por rechazar una expansión de la mente de estas características. Al respecto yo señalaría que esta inteligencia adicional residirá fundamentalmente en la nube (la red de ordenadores en aumento exponencial a la que nos conectamos a través de las comunicaciones online), donde la mayor parte de la inteligencia de

nuestras máquinas está ya almacenada. Cuando utilizamos un buscador, un sistema de reconocimiento del habla en nuestro teléfono móvil, consultamos un asistente virtual como Siri o usamos nuestro teléfono móvil para traducir un signo en otra lengua, la inteligencia no reside en el propio dispositivo, sino en la nube. Ahí también se hospedará nuestro neocórtex expandido. El que accedamos a dicha inteligencia expandida a través de la conexión neuronal directa o de la forma en que lo hacemos ahora (interaccionando con ella por medio de nuestros dispositivos) es una distinción arbitraria. En mi opinión, todos nos volveremos más creativos gracias a esta mejora generalizada, independientemente de que optemos o rechacemos una conexión directa a la inteligencia expandida de la humanidad. Ya hemos externalizado en la nube gran parte de nuestra memoria personal, social, histórica y cultural, y en último término haremos lo mismo con nuestro pensamiento jerárquico.

El logro de Einstein no solo fue el resultado de su aplicación de metáforas por medio de experimentos mentales, sino también de su valentía al creer en el poder de dichas metáforas. Estuvo dispuesto a renunciar a las explicaciones tradicionales que habían fracasado a la hora de satisfacer sus experimentos, y asimismo estuvo dispuesto a soportar el ridículo al que sus coetáneos condenaban a las extrañas explicaciones que implicaban sus metáforas. Estas cualidades (confianza en la metáfora y coraje en la convicción) son las que también deberíamos ser capaces de programar en nuestro neocórtex no biológico.

## *Amor*

> Claridad de mente también significa claridad de pasión. Por eso una mente grande y clara ama ardientemente y percibe claramente lo que ama.
>
> —BLAISE PASCAL

> En el amor siempre hay algo de locura. Sin embargo, siempre hay algo de razón en la locura.
>
> —FRIEDRICH NIETZSCHE

> Cuando hayas visto tanto de la vida como yo, no subestimarás el poder del amor obsesivo.
>
> —ALBUS DUMBLEDORE, EN *HARRY POTTER Y EL MISTERIO DEL PRÍNCIPE* (J. K. ROWLING)

> Siempre disfruto de una buena solución matemática para cualquier problema amoroso.
>
> —MICHAEL PATRICK KING, EN “TAKE ME OUT TO THE BALLGAME”, EPISODIO DE *SEX AND THE CITY*

Si usted no ha experimentado personalmente el amor extático, desde luego que ha oído hablar de él. Es justo afirmar que una parte importante, si no la mayor parte del

arte, ya se trate de historias, novelas, música, danza, pintura, espectáculos televisivos o películas, está inspirada en historias de amor incipiente.

Recientemente, también la ciencia se ha visto involucrada y ya somos capaces de identificar los cambios bioquímicos que se producen cuando alguien se enamora. Se segrega dopamina, lo que produce sentimientos de felicidad y deleite. Los niveles de noradrenalina se disparan, lo cual hace que se acelere el corazón y que se experimenten sentimientos de júbilo. Estas sustancias químicas, junto a la feniletilamina, producen euforia, altos niveles de energía, concentración, pérdida del apetito y, por lo general, ansias por el objeto de deseo. Curiosamente, una investigación reciente del University College de Londres también demuestra que los niveles de serotonina bajan, algo similar a lo que ocurre con los desórdenes obsesivo-compulsivos. Esto es congruente con la naturaleza obsesiva del amor incipiente (2). Los altos niveles de dopamina y la noradrenalina explican el aumento de la atención a corto plazo, la euforia y las ansias de este tipo de amor.

Si estos fenómenos bioquímicos resultan parecidos a los del síndrome de lucha o huida es porque lo son, si exceptuamos que aquí lo que hacemos es correr hacia algo o alguien (de hecho, un cínico podría decir que en este caso se corre hacia el peligro en vez de alejarse de él). Estos cambios también coinciden plenamente con los de las primeras fases de los comportamientos adictivos. La canción de Roxy Music "Love Is the Drug" es bastante precisa describiendo este estado (no obstante, el tema de la canción es la búsqueda de su siguiente dosis de amor). Estudios sobre las experiencias de éxtasis religioso también muestran los mismos fenómenos físicos. Así, se podría decir que la persona que tiene una de estas experiencias se está enamorando de Dios o de cualquiera que sea la conexión espiritual en la que esté centrada la experiencia.

En el caso del amor romántico incipiente, el estrógeno y la testosterona ciertamente juegan un importante papel a la hora de establecer la atracción sexual, pero si la reproducción sexual fuera el único objetivo evolutivo del amor, entonces el aspecto romántico del proceso sería innecesario. Tal y como escribió el psicólogo William Money (1921–2006) "el deseo es lascivo, el amor es lírico".

La fase extática del amor conduce hasta la fase del cariño y en último término a una unión a largo plazo. También existen productos químicos que favorecen este proceso, incluyendo la oxitocina y la argipresina. Tomemos dos especies cercanas de ratones de campo: el microtus ochrogaster y el microtus montanus. Son prácticamente idénticos, exceptuando el hecho de que el primero posee receptores de oxitocina y argipresina, mientras que el segundo no los tiene. El primero se caracteriza por tener relaciones monógamas durante toda la vida, mientras que el segundo recurre casi exclusivamente a relaciones de una noche. Por lo tanto podemos decir que, en el caso de los ratones de campo, los receptores de oxitocina y argipresina son en gran medida los que determinan la naturaleza de sus vidas

amorosas.

Aunque estos productos químicos también influyen en los humanos, nuestro neocórtex ha asumido el papel dominante, al igual que lo ha hecho en todas las demás facetas. Los ratones de campo tienen neocórtex, pero es del tamaño de un sello postal, plano y lo suficientemente grande como para permitirles encontrar un compañero para el resto de su vida (o, en el caso del microtus montanus, por lo menos para una noche), así como para desempeñar los comportamientos básicos de los ratones de campo. Los humanos tenemos el neocórtex lo suficientemente desarrollado como para participar en la amplia gama de expresiones "líricas" a las que se refiere Money.

Desde una perspectiva evolutiva, el propio amor existe para cumplir con las necesidades del neocórtex. Si no tuviéramos un neocórtex, el deseo bastaría para garantizar la reproducción. No obstante, la instigación extática del amor nos lleva al cariño y al amor maduro, y da como resultado una unión duradera. A su vez, esto está diseñado así para por lo menos proporcionar la posibilidad de un entorno estable para los niños cuando sus neocórtex experimenten el fundamental proceso de aprendizaje que les convierte en adultos responsables y competentes. El aprendizaje en un ambiente fértil forma parte inherente del método que sigue el neocórtex. De hecho, los mismos mecanismos hormonales de la oxitocina y de la argipresina juegan un papel fundamental en el crucial establecimiento de relaciones entre padres (sobre todo la madre) e hijos.

El amor llevado a su extremo hace que el ser querido se convierta en una parte importante de nuestro neocórtex. Después de décadas juntos, en el neocórtex existe un otro virtual, de manera que podemos anticipar cada paso o palabra que nuestro ser querido tome o diga. Nuestros patrones neocorticales están llenos de pensamientos y patrones que reflejan quienes son. Cuando perdemos a un ser querido, literalmente perdemos parte de nosotros mismos. No se trata *solamente* de una metáfora, todos los extensos reconocedores de patrones llenos de los patrones que reflejan a la persona amada de repente cambian de naturaleza. Aunque pueden ser considerados como una valiosa manera de mantener viva a dicha persona en nuestro interior, los extensos patrones neocorticales correspondientes a alguien que hemos perdido cambian súbitamente de detonantes de placer a detonantes de luto.

La base evolutiva del amor y sus correspondientes fases no explican por completo lo que ocurre en el mundo de hoy. En gran medida ya hemos conseguido liberar al sexo de su función biológica, ya que podemos tener hijos sin recurrir al sexo y podemos tener relaciones sexuales sin tener que engendrar hijos. En su mayor parte, mantenemos relaciones sexuales por razones sensuales y relacionales. Asimismo, permanentemente nos enamoramos por razones que no son la de tener descendencia.

De igual forma, el enorme campo de la expresión artística que se dedica a celebrar el amor en sus innumerables formas se remonta hasta la antigüedad y también constituye un fin en sí mismo. Nuestra capacidad para crear estas imperecederas formas de conocimiento transcendente, bien en lo que se refiere al amor o a cualquier otra cosa, es precisamente lo que hace única a nuestra especie.

El neocórtex es la máxima creación de la biología. A su vez, los poemas sobre el amor y todo el resto de nuestras creaciones representan las invenciones más importantes de nuestro neocórtex.

* Traducción de http://pdcrodas.webs.ull.es/literatura/shakespeare_sonetos.htm#a73: Ese tiempo del año puedes en mí contemplar cuando hojas amarillas, o ninguna, o pocas, cuelgan de esas ramas que tiemblan contra el frío, desnudos coros arruinados donde recientemente cantaban los dulces pájaros: en mí ves el crepúsculo del día que tras el ocaso se va apagando en el poniente, el cual poco a poco la negra noche se lleva, segundo yo de la muerte que todo lo sella en el descanso: en mí ves la lumbre del fuego que sobre las cenizas de su juventud reposa como el lecho de muerte sobre el que debe expirar, consumido por lo que lo nutrió: esto percibes, lo que te hace el amor más fuerte para amar bien lo que has de dejar en breve.

CAPÍTULO SIETE

# El neocórtex digital de inspiración biológica

No confíes nunca en algo que pueda pensar por sí mismo si no puedes ver dónde tiene el cerebro.

—ARTHUR WEASLEY, EN J. K. ROWLING, *HARRY POTTER Y EL PRISIONERO DE AZKABAN*

No, no estoy interesado en desarrollar un cerebro poderoso. A todo lo que aspiro es a un cerebro mediocre como el del Presidentes de *American Telegraph Company and Telegraph Company*.

—ALAN TURING

Un ordenador merecería ser tildado de inteligente si pudiera engañar a una persona y hacerle creer que es humano.

—ALAN TURING

Creo que al final de siglo el uso de las palabras y la opinión generalizada entre las personas cultas habrán cambiado tanto que será posible referirse a máquinas pensantes sin esperar ser contradicho.

—ALAN TURING

Una rata madre es capaz de construir un nido para sus crías aunque no haya visto otra rata en su vida (1). De forma similar, una araña es capaz de tejer una tela de araña, una oruga es capaz de crear su propio capullo y un castor es capaz de crear una presa incluso sin que ningún coetáneo les haya enseñado cómo realizar estas tareas tan complejas. Esto no significa que no existan los comportamientos aprendidos, significa simplemente que estos animales no aprendieron dichos comportamientos durante una sola vida, sino que los aprendieron durante miles de vidas. La evolución del comportamiento animal se conforma en un proceso de aprendizaje. Sin embargo, se trata de un aprendizaje por parte de la especie, no del individuo, y los frutos de este proceso de aprendizaje se codifican en el ADN.

Para apreciar el significado de la evolución del neocórtex hay que tener en cuenta que aceleró enormemente el proceso de aprendizaje (el conocimiento jerárquico), que pasó de ser de miles de años a ser de meses (o incluso menos). Incluso si millones de animales pertenecientes a una especie de mamíferos en particular fracasaran a la hora de resolver un problema que requiriera de una

jerarquía de pasos, solo sería necesario que uno de los individuos diera con la solución de forma accidental. El nuevo método sería copiado y se extendería de forma exponencial por toda la población.

En estos momentos, el paso de la inteligencia biológica a la no biológica nos permite volver a acelerar el proceso de aprendizaje miles o millones de veces. Una vez que un neocórtex digital adquiere una capacidad, puede transferir dicho conocimiento en cuestión de minutos o incluso de segundos. Valga como un ejemplo de entre los muchos que cabría citar el de mi primera empresa, Kurzweil Computer Products (hoy Nuance Speech Technologies), que fundé en 1973. Nos pasamos años entrenando un conjunto de ordenadores orientados a la investigación para que reconocieran las letras impresas de documentos escaneados, una tecnología llamada reconocimiento óptico de caracteres omni-font (cualquier tipo de fuente), cuyo acrónimo en inglés es OCR. Esta tecnología en concreto ha sufrido un desarrollo constante durante casi cuarenta años y su fruto hoy se llama OmniPage, un producto de Nuance. A diferencia de lo que nosotros hacíamos, si usted quiere que su ordenador reconozca letras impresas, no tiene que pasarse años entrenándole, no tiene más que descargarse, bajo la forma de software, la evolución de patrones previamente aprendida por los ordenadores destinados a la investigación. En la década de 1980, comenzamos con el reconocimiento del habla. Dicha tecnología, que también se ha visto sujeta a un desarrollo constante durante varias décadas, forma parte de Siri. Así, usted puede, en cuestión de segundos, descargarse la evolución de los patrones aprendida por los ordenadores destinados a la investigación durante muchos años.

En último término, crearemos un neocórtex artificial que posea toda la gama y flexibilidad de su homólogo humano. Consideremos los beneficios de esto. Los circuitos electrónicos son millones de veces más rápidos que nuestros circuitos biológicos. Al principio, tendremos que dedicar todo este aumento de velocidad a compensar la relativa falta de paralelismo de nuestros ordenadores, pero en último término el neocórtex digital será mucho más veloz que el biológico y su velocidad no hará más que aumentar.

Cuando acrecentemos nuestro neocórtex mediante una versión sintética, no tendremos que preocuparnos sobre qué cantidad de neocórtex adicional puede caber físicamente en nuestros cuerpos y cerebros, ya que la mayor parte estará en la nube, al igual que la mayor parte de la computación que usamos hoy en día. Antes calculé que, incluso teniendo en cuenta la innovación evolutiva que supone tener una frente amplia y el hecho de que el neocórtex ocupe alrededor del 80% del espacio disponible, en nuestro neocórtex biológico poseemos del orden de 300 millones de reconocedores de patrones. Eso es lo máximo que podríamos meter en el interior de nuestros cráneos. A partir del momento en que empecemos a pensar en la nube los límites naturales dejarán de existir, seremos capaces de usar miles de millones o

billones de reconocedores de patrones según sean nuestras necesidades y según nos lo permita en cada momento la ley de los rendimientos acelerados.

Para hacer que un neocórtex digital aprenda una nueva capacidad, seguirán siendo necesarias muchas iteraciones educativas, tal y como ocurre con el cerebro biológico, pero una vez que un solo neocórtex digital, en alguna parte y en un cierto momento, aprenda algo, podrá compartirlo inmediatamente con todo el resto de neocórtex digitales. En la nube podremos tener nuestros expansores de neocórtex privados, igual que ahora tenemos nuestros almacenamientos de datos personales.

Por último, pero no por ello menos importante, seremos capaces de realizar copias de seguridad de la parte digital de nuestra inteligencia. Tal y como hemos visto, el decir que nuestro neocórtex contiene información no es solo una metáfora y da miedo ver cómo hoy en día nada de esa información puede almacenarse en copias de seguridad. Por supuesto, existe una manera en la que sí que podemos realizar copias de seguridad de la información contenida por nuestros cerebros: escribiendo dicha información. La capacidad de transferir, al menos en parte, nuestro pensamiento en un sustrato que puede sobrevivir a nuestros cuerpos biológicos supuso un gran paso adelante, pero una gran cantidad de datos en nuestro cerebro sigue siendo vulnerable.

## *Simulaciones cerebrales*

Una estrategia para construir un cerebro digital es la de simular de forma precisa un cerebro biológico. Por ejemplo, David Dalrymple, nacido en 1991 y estudiante de doctorado en ciencias del cerebro en Harvard, está planeando simular el cerebro de un nematodo, en concreto de un ascáride (2). Dalrymple optó por este nematodo debido a su sistema nervioso relativamente simple, el cual consta de alrededor de 300 neuronas. El investigador planea simularlo a nivel molecular de forma muy exhaustiva. También creará una simulación informática del cuerpo del animal, así como de su entorno, de manera que el nematodo virtual pueda cazar comida (virtual) y realizar el resto de cosas que los nematodos suelen hacer. Dalrymple dice que es probable que esto sea la primera carga cerebral completa desde un animal biológico a otro virtual que vive en un mundo virtual. Al igual que ocurre con su nematodo simulado, la cuestión sobre si los nematodos biológicos tienen o no consciencia sigue siendo cuestión de debate, aunque durante sus esfuerzos por conseguir comida, digerirla, evitar depredadores y reproducirse sí que tienen experiencias de las que son conscientes.

En el extremo opuesto del espectro, el Blue Brain Project de Henry Markram planea simular el cerebro humano, incluido todo el neocórtex así como regiones del

cerebro antiguo tales como el hipocampo, la amígdala y el cerebelo. Las simulaciones planeadas se construirán según diferentes grados de precisión hasta llegar a una simulación completa a nivel molecular. Tal y como señalé en el capítulo 4, Markram ha descubierto un módulo fundamental con varias docenas de neuronas que se repite en el neocórtex una y otra vez, lo cual demuestra que el aprendizaje se realiza mediante estos módulos y no mediante neuronas individuales.

Los progresos hechos por Markram han aumentado de forma exponencial. En 2005, año en el que se inauguró el proyecto, simuló una neurona. En 2008 su equipo simuló una columna neocortical completa, compuesta por 10 000 neuronas, perteneciente al cerebro de una rata. En 2011 la cifra aumentó hasta las 100 columnas que hacen un total de un millón de células a las que denomina como un mesocircuito. Un punto controvertido del trabajo de Markram es cómo verificar si las simulaciones son precisas. Para ello estas simulaciones tendrán que demostrar que aprenden tal y como expongo más abajo.

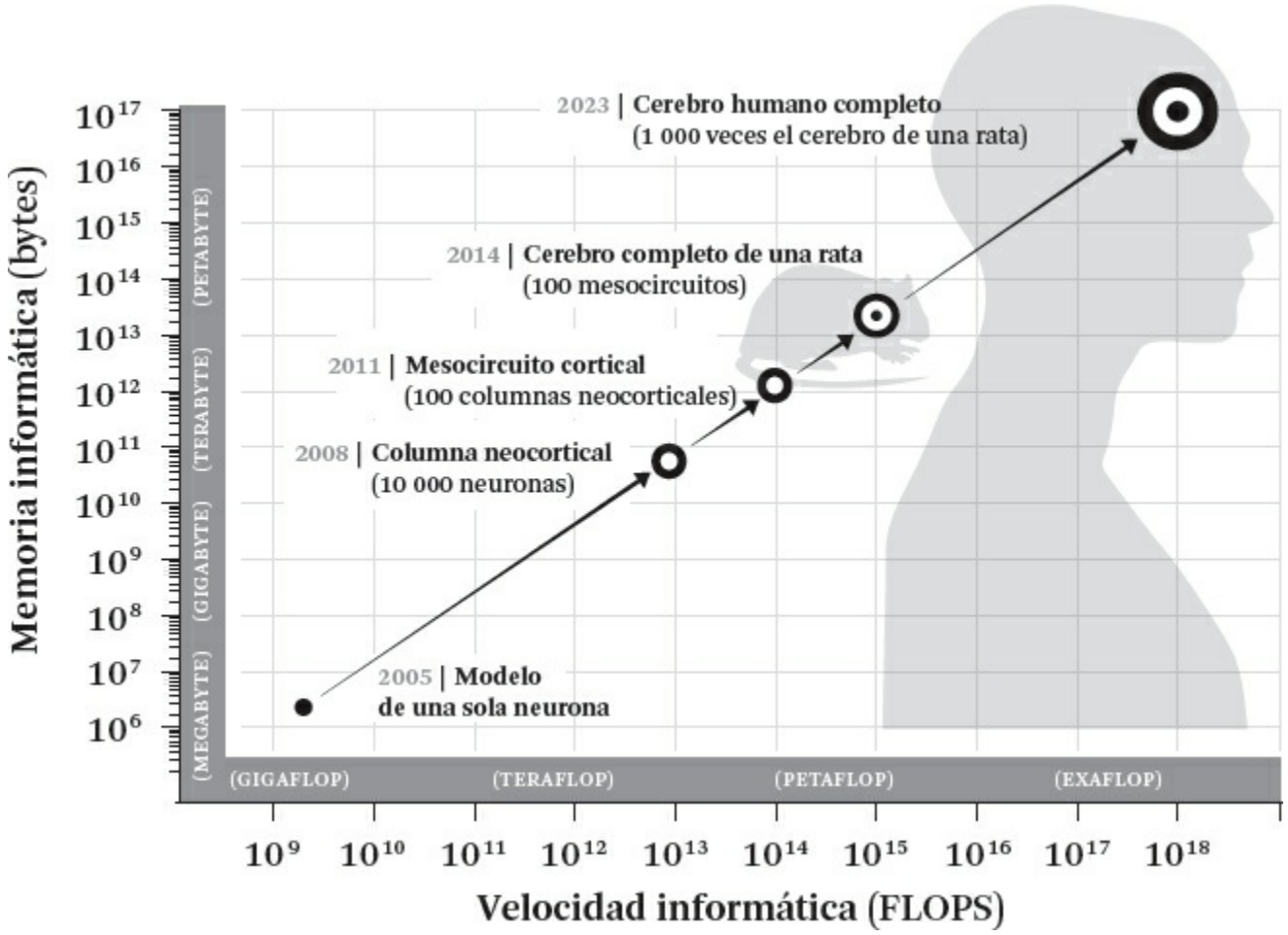
Markram planea simular el cerebro completo de una rata y sus 100 mesocircuitos, un total de 100 millones de neuronas y aproximadamente un billón de sinapsis, en el año 2014. Durante una conferencia en el TED de Oxford de 2009, Markram dijo: “no es imposible construir un cerebro humano, lo podemos conseguir en 10 años” (3). Su previsión más reciente para logar una simulación cerebral completa es para el año 2023 (4).

Markram y su equipo fundamentan su modelo en detallados análisis anatómicos y electromecánicos de neuronas reales. Mediante un dispositivo automatizado han creado un robot de fijación de voltaje* para medir canales iónicos específicos, neurotransmisores y enzimas responsables de la actividad electromecánica en el interior de cada neurona. Su sistema automatizado fue capaz de realizar 30 años de análisis en solo seis meses, según Markram. Fue a partir de estos análisis por lo que se dieron cuenta de la existencia de las unidades de “memoria Lego” que representan las unidades funcionales básicas del neocórtex.

Contribuciones importantes a la tecnología de la fijación de voltaje robótica fueron hechas por el neurocientífico del MIT llamado Ed Boyden, por el profesor de ingeniería mecánica de Georgia Tech llamado Craig Forest y por el estudiante de doctorado del departamento de Forest llamado Suhasa Kodandaramaiah. Estas personas hicieron público un sistema automatizado cuya precisión es de un micrómetro y que puede escanear tejido neuronal a distancias muy pequeñas sin dañar las delicadas membranas de las neuronas. “Se trata de algo que un robot puede hacer, pero un humano no”, comentó Boyden.

Volviendo a la simulación de Markram, después de que simulara una columna neocortical, a Markram se le atribuyó la siguiente declaración: “ahora solo tenemos que ampliarlo” (5). La ampliación es sin duda un factor importante, pero existe una barrera fundamental: el aprendizaje. Si el cerebro del proyecto Brain Project está

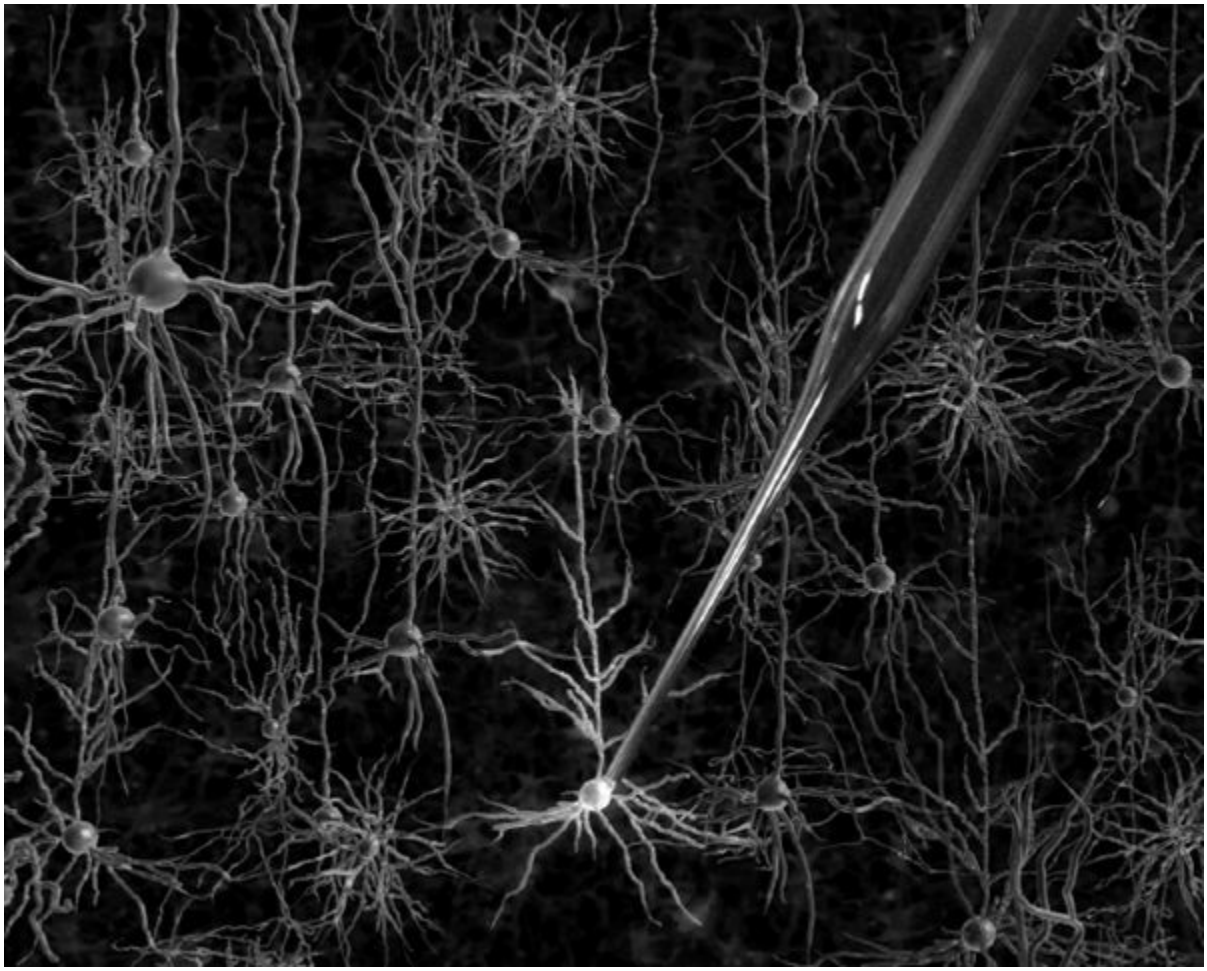
llamado a “hablar y poseer un inteligencia y un comportamiento muy parecidos a los humanos”, que es como Markram describió su objetivo en una entrevista con la BBC en 2009, entonces necesitará disponer del contenido necesario en el interior del neocórtex simulado como para llevar a cabo dichas tareas (6). Tal y como cualquiera que haya intentado mantener una conversación con un recién nacido puede atestiguar, antes de que algo así pueda conseguirse se necesita mucho aprendizaje.




**Progreso actual y proyectado del proyecto de simulación cerebral Blue Brain.**

Existen dos maneras obvias en las que esto puede conseguirse en un cerebro simulado como Blue Brain. Una sería hacer que el cerebro aprendiera estos contenidos de la manera en que lo hace un cerebro humano. Puede empezar como si fuera un recién nacido humano dotado de una capacidad innata para adquirir conocimientos jerárquicos y dotado de ciertas transformaciones preprogramadas en sus regiones sensoriales de preprocesamiento. Sin embargo, el aprendizaje que separa a un niño biológico de una persona que puede mantener una conversación tendría que producirse en el aprendizaje no biológico de una manera similar a la que se produce en el aprendizaje biológico. El problema de esta estrategia es que un cerebro que esté siendo simulado al nivel de detalle anticipado por Blue Brain no es de esperar que funcione en tiempo real hasta por lo menos el principio de la década de 2020. Aun así el funcionamiento en tiempo real sería demasiado lento a no ser

que los investigadores estén dispuestos a esperar una década o dos hasta alcanzar la paridad intelectual con respecto a un humano adulto, pese al hecho de que el rendimiento a tiempo real se volverá constantemente más veloz a medida que los ordenadores continúen mejorando su relación rendimiento/precio.



**La punta del robot de fijación de voltaje desarrollado en el MIT y en Georgia Tech escaneando tejido neuronal.**

La otra estrategia es tomar uno o más cerebros biológicos humanos que ya hayan adquirido los suficientes conocimientos como para conversar utilizando un lenguaje dotado de significado y además se comporten de forma madura, para después copiar sus patrones neocorticales en el interior del cerebro simulado. El problema con este método es que requiere de una tecnología de escaneado no invasiva y no destructiva dotada de la suficiente resolución espacial y temporal y con la suficiente velocidad como para realizar dicha tarea rápida y completamente. No obstante, no preveo que una tecnología "de carga" así esté disponible hasta más o menos la década de 2040. (Las necesidades computacionales para simular un cerebro a ese nivel de precisión, que calculo que es de $10^{19}$ cálculos por segundo, estará disponible en un superordenador que según mis proyecciones datará de principios de la década de 2020. Sin embargo, las necesarias tecnologías de escaneado cerebral no destructivas tardarán más tiempo en aparecer).

Existe una tercera estrategia, que es la que creo que los proyectos de simulación tales como Blue Brain tendrán que poner en práctica. Los modelos moleculares pueden simplificarse creando equivalentes funcionales a diferentes niveles de especificidad que van desde mi propio método algorítmico (tal y como se describe en este libro) hasta las simulaciones más cercanas a las simulaciones moleculares completas. Así, la velocidad de aprendizaje puede ser incrementada en un factor de cientos de miles de veces dependiendo del grado de simplificación utilizado. Mediante el modelo funcional se puede idear un programa de formación que el cerebro simulado pueda aprender relativamente rápido. Entonces la simulación molecular completa podría ser sustituida por el modelo simplificado y seguir utilizándose su aprendizaje acumulado. Así podríamos simular el aprendizaje mediante el modelo molecular completo a una velocidad mucho más baja.

El informático norteamericano Dharmendra Modha y sus colegas de IBM han creado una simulación célula-a-célula de una parte del neocórtex visual humano que contiene 1,6 miles de millones de neuronas virtuales y 9 billones de sinapsis, lo cual equivale al neocórtex de un gato. En un superordenador IBM BlueGene/P, que consta de 147 456 procesadores, funciona 100 veces más despacio que el tiempo real. Este trabajo recibió el premio Gordon Bell de la *Association for Computing Machinery*.

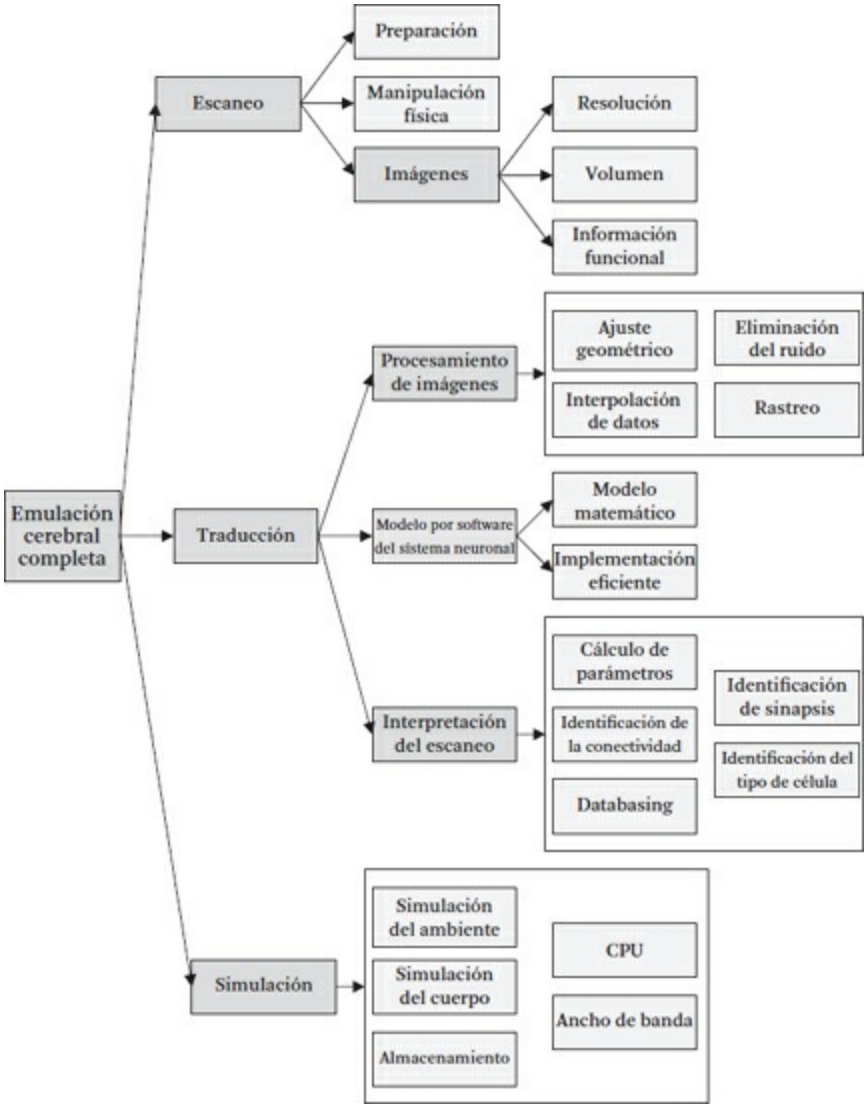
El objetivo de un proyecto de simulación cerebral como Blue Brain y de las simulaciones del neocórtex de Modha es específicamente el de perfeccionar y demostrar un modelo funcional. La inteligencia artificial de nivel humano usará sobre todo el tipo de modelo algorítmico expuesto en este libro. Sin embargo, las simulaciones moleculares nos ayudarán a perfeccionar dicho modelo y a comprender por completo qué detalles son importantes. Durante mi desarrollo de la tecnología de reconocimiento del habla en las décadas de 1980 y 1990, fuimos capaces de perfeccionar nuestros algoritmos una vez que las transformaciones reales llevadas a cabo por el nervio auditivo y capas iniciales del córtex auditivo fueron entendidas. Incluso si nuestro modelo funcional fuera perfecto, llegar a entender exactamente cómo se implementa realmente en nuestros cerebros biológicos revelará datos importantes sobre la función y disfunción humana.

Necesitaremos datos precisos sobre cerebros reales que nos permitan crear simulaciones basadas en la biología. Por eso el equipo de Makram está recogiendo sus propios datos. Existen proyectos a gran escala cuyo objetivo es reunir este tipo de datos y ponerlos a disposición de los científicos en general. Por ejemplo, *Cold Spring Harbor Laboratory* de Nueva York ha reunido 500 terabytes de datos procedentes del escaneo del cerebro de un mamífero (un ratón) y en junio de 2012 hicieron públicos estos datos. Su proyecto le permite al usuario explorar un cerebro de forma similar a como Google Earth permite explorar la superficie del planeta. Permite moverse por todo el cerebro y ampliarlo para poder ver las neuronas

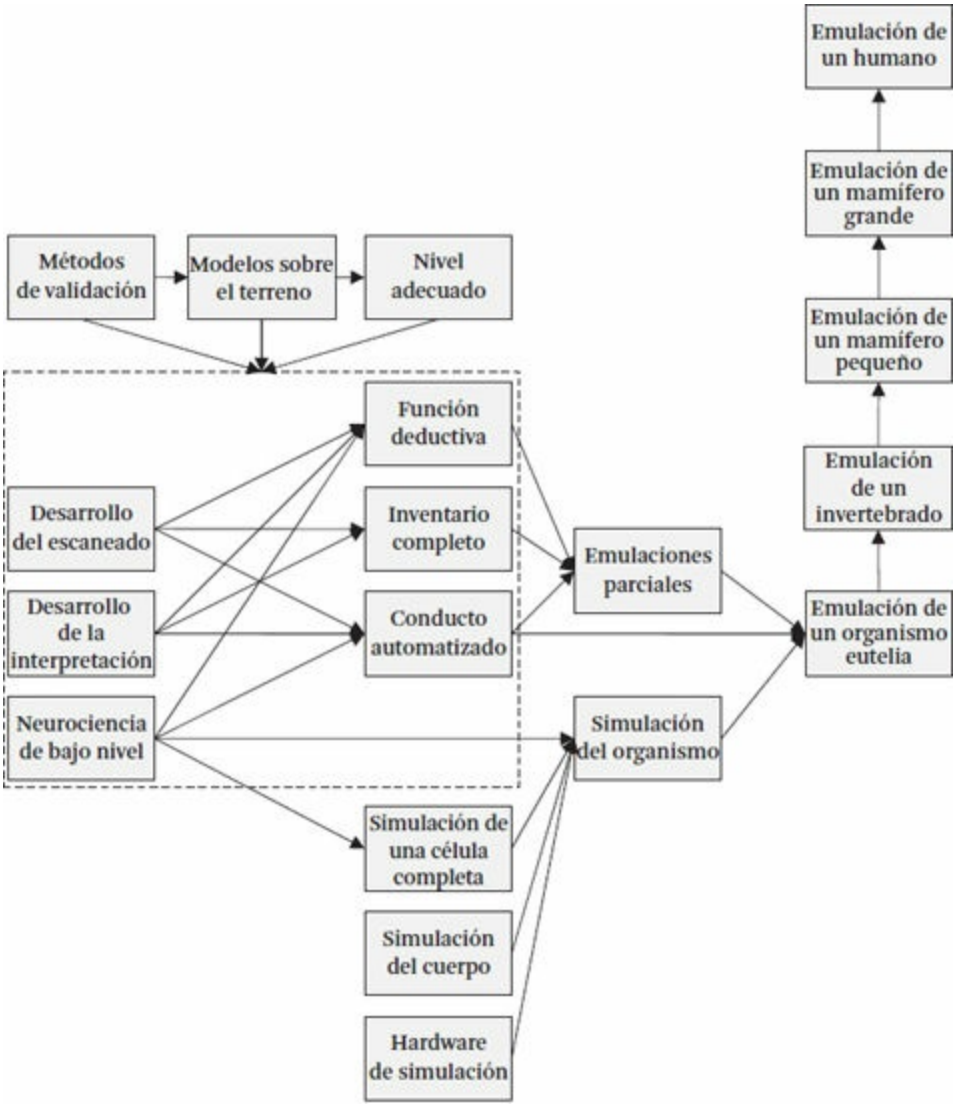
individuales y sus conexiones. Además, permite resaltar una conexión concreta y luego seguir su rastro a lo largo del cerebro.

Dieciséis secciones del los Institutos Nacionales de Salud se han reunido para financiar una iniciativa a gran escala llamada *Human Connectome Project*, dotada de un presupuesto de 38,5 millones de dólares (7). Bajo la dirección de la *Washington University in St. Louis, University of Minnesota, Harvard University, Massachusetts General Hospital* y *University of California at Los Angeles*, el proyecto pretende crear un mapa tridimensional similar sobre las conexiones del cerebro humano. El proyecto está utilizando varias tecnologías de escaneo no invasivas, incluyendo nuevas formas de MRI, magnetoencefalografía (la medición de los campos magnéticos producidos por la actividad cerebral del cerebro) y la tractografía por difusión (un método para seguir el rastro de los fibrados del cerebro). Tal y como indico en el capítulo 10, la resolución espacial del escaneo no invasivo del cerebro está mejorando de forma exponencial. La investigación llevada a cabo por Van J. Wedeen y sus colegas del *Massachusetts General Hospital* muestra una estructura en red del cableado del cerebro altamente regular a la que me he referido en el capítulo 4, lo cual supone uno de los primeros éxitos de este proyecto.

Anders Sandberg, el neurocientífico computacional de la Universidad de Oxford nacido en 1972, y el filósofo sueco Nick Bostrom (nacido en 1973) han escrito la exhaustiva obra *Whole Brain Emulation: A Roadmap*, que detalla los requisitos para simular el cerebro humano (y otros tipos de cerebros) a diferentes niveles de especificidad, desde modelos funcionales de alto nivel hasta la simulación de moléculas (8). El informe no proporciona ninguna estimación temporal, pero describe los requisitos para simular los diferentes tipos de cerebros a diferentes niveles de precisión en términos de escaneo cerebral, modelización, almacenamiento y computación. El informe prevé mejoras exponenciales que ya se están produciendo en todas capacidades de estas áreas y sostiene que los requisitos para simular el cerebro humano a un nivel elevado de detalle ya están siendo generados.




**Resumen de las capacidades tecnológicas necesarias para una emulación cerebral completa, perteneciente a *Whole Brain Emulation: A Roadmap* (de Anders Sandberg y Nick Bostrom).**




**Esquema de *Whole Brain Emulation: A Roadmap* (de Anders Sandberg y Nick Bostrom).**

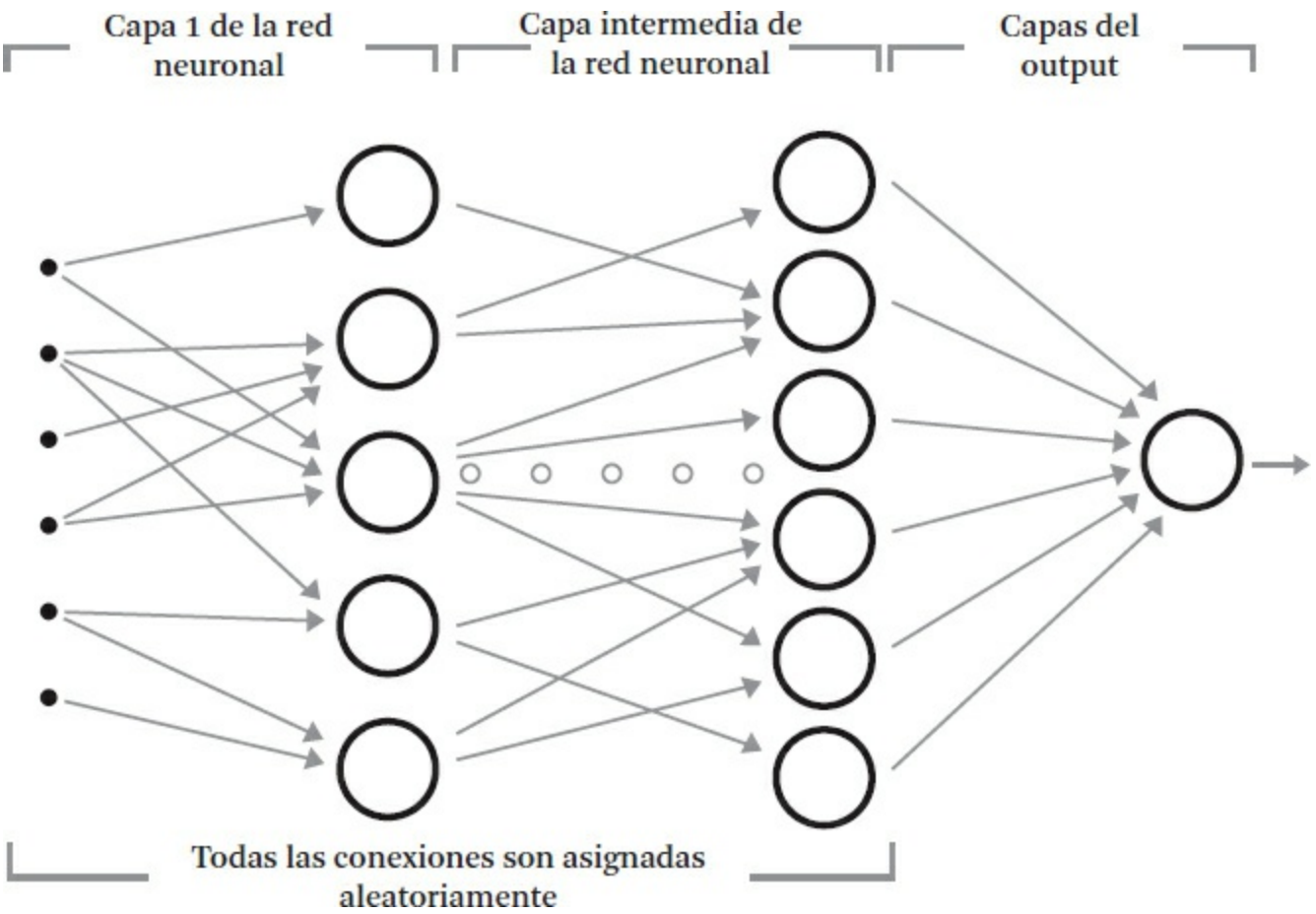
## *Redes neuronales*

En 1964, con 16 años, escribí una carta a Frank Rosenblatt (1928–1971), profesor en la Universidad Cornell, interesándome sobre una máquina llamada Mark 1 Perceptron. Él la había creado hacía 4 años y era descrita como poseedora de propiedades similares a las del cerebro. Me invitó a visitarle y a que probara la máquina.

El Perceptron había sido construido a partir de lo que él decía que eran modelos electrónicos de neuronas. El input consistía en valores ordenados en dos dimensiones. Para el habla, una dimensión representaba la frecuencia y otra el tiempo, de manera que cada valor representaba la intensidad de una frecuencia en un determinado punto temporal. Para las imágenes, cada punto era un pixel de una imagen bidimensional. Cada punto de un input determinado era conectado aleatoriamente a los inputs de la primera capa de neuronas simuladas. Cada conexión

tenía asociada una potencia sináptica que representaba su importancia y que inicialmente estaba fijada en un valor aleatorio. Cada neurona sumaba las señales que le llegaban. Si la señal combinada superaba un determinado umbral, la neurona se disparaba y mandaba una señal a su conexión de output; si la señal de input combinada no superaba el umbral la neurona no se disparaba y su output era cero. El output de cada neurona estaba aleatoriamente conectado a los inputs de las neuronas de la capa siguiente. El Mark 1 Perceptron tenía tres capas que podían organizarse según diferentes configuraciones. Por ejemplo, una capa podía retroalimentar otra capa anterior. En la capa más alta, el output de una o más neuronas (también seleccionadas aleatoriamente) proporcionaba la contestación. (Para una descripción algorítmica de las redes neuronales, véase la nota (9)).

Dado que inicialmente el cableado de la red neuronal y los pesos sinápticos son fijados de forma aleatoria, las contestaciones de una red neuronal no entrenada también son aleatorias. Por tanto, la clave de una red neuronal es que debe aprender su materia de estudio, igual que los cerebros de los mamíferos sobre los que se supone que está inspirada. Una red neuronal parte desde la ignorancia. Su profesor, que puede ser un ser humano, un programa informático o quizás otra red neuronal más madura que ya se haya aprendido las lecciones, recompensa a la red neuronal estudiante cuando esta genera el output correcto y la castiga cuando no lo hace. A su vez, este *feedback* es utilizado por la red neuronal estudiante para ajustar la potencia de cada conexión interneuronal. Las conexiones que sean congruentes con la respuesta correcta se refuerzan. Las que optan por una respuesta errónea son debilitadas.




Con el tiempo, la red neuronal se autoorganiza para proporcionar las respuestas correctas sin necesidad de ayuda. Experimentos han demostrado que las redes neuronales pueden aprender su objeto de estudio teniendo incluso profesores poco fiables. Si el profesor tiene razón tan solo el 60% de las veces, la red neuronal estudiante seguirá aprendiéndose la lecciones con una precisión cercana al 100%.

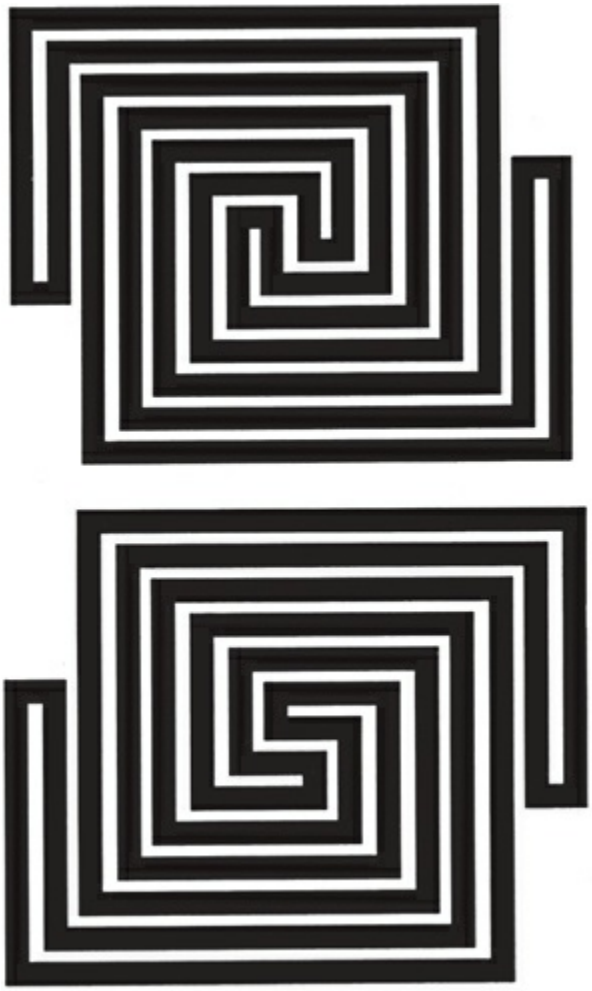
Sin embargo, las limitaciones en la gama de materias en las que el Perceptron era capaz de aprender se hicieron rápidamente patentes. Cuando visité al profesor Rosenblatt en 1964 probé a hacer sencillas modificaciones en el input. El sistema estaba configurado para reconocer letras impresas y de hecho las reconocía de forma bastante precisa. También era bastante bueno en el campo de la autoasociación (es decir, que podía reconocer las letras incluso si tapaba partes de ellas), pero era bastante peor en el campo de la invarianza (es decir, en la generalización sobre cambios de tamaño y fuente, las cuales confundía).

Durante la segunda mitad de la década de 1960, estas redes neuronales se volvieron enormemente populares y el campo del “conexionismo” llegó a significar por lo menos la mitad del campo de la inteligencia artificial. Mientras tanto, la estrategia de la IA más tradicional incluía intentos directos por programar soluciones a problemas específicos, como por ejemplo cómo reconocer las propiedades invariables de las letras impresas.

Otra persona a la que visité en 1964 fue Marvin Minsky (nacido en 1927), uno de

los fundadores del campo de la inteligencia artificial. A pesar de que él mismo había hecho en la década de 1950 ciertos trabajos pioneros en el campo de las redes neuronales, se mostró preocupado por la gran explosión de interés en esta técnica. Parte de la fascinación por las redes neuronales provenía del hecho de que se suponía que no necesitaban de programación, sino que aprendían las soluciones a los problemas por sí solas. En 1965 ingresé como estudiante en el MIT con el profesor Minsky como mentor y compartí su escepticismo sobre la moda del "conexionismo".

En 1969, Minsky y Seymour Papert (nacido en 1928), los dos cofundadores del Laboratorio de Inteligencia Artificial del MIT, escribieron un libro llamado *Perceptrons* que exponía un solo teorema central: el hecho de que un perceptron era intrínsecamente incapaz de determinar si una imagen estaba o no conectada. El libro desató una tormenta de fuego. Determinar si una imagen está o no conectada es una tarea que los humanos pueden hacer muy fácilmente y el programar un ordenador para que haga esta discriminación también es un proceso sencillo. El hecho de que el Perceptron no pudiera hacer esto fue considerado por muchos como un defecto fatal.



**Dos imágenes de la portada del libro *Perceptrons*, de Marvin Minsky y Seymour Papert. La imagen de arriba no está conectada (es decir, el área oscura consta de dos partes desconectadas). La imagen de abajo está conectada. Un humano puede determinar esto fácilmente, al igual que un sencillo programa de software. Un perceptron de prealimentación como el Mark 1 Perceptron de Frank Rosenblatt no puede hacer esta discriminación.**

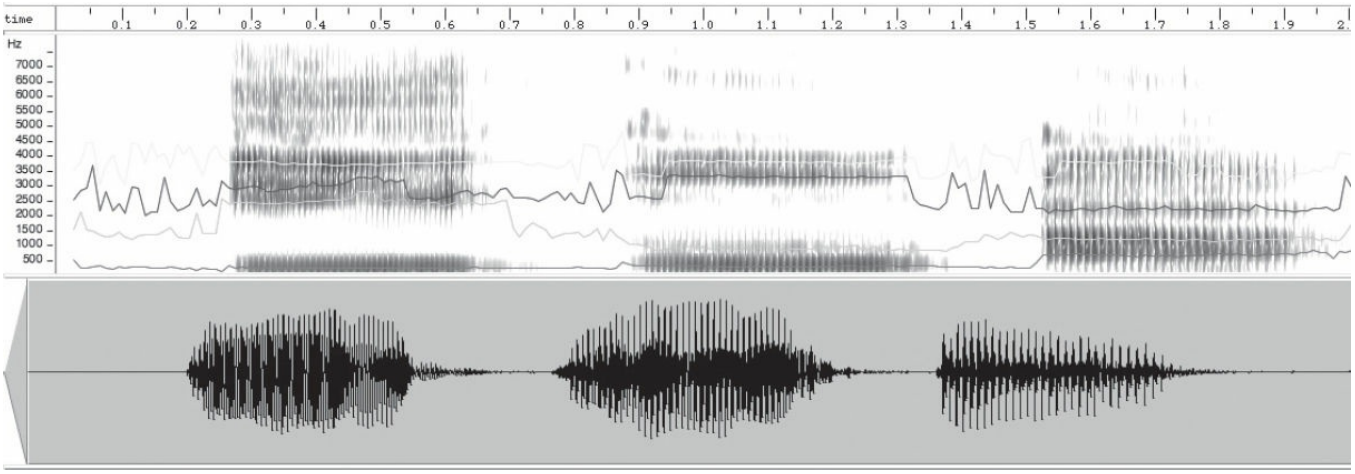
Sin embargo, la opinión mayoritaria sobre el Perceptron era que este podía hacer más cosas de las que en realidad era capaz de hacer. El teorema de Minsky y Papert solo era aplicable a un tipo concreto de red neuronal llamada red neuronal de prealimentación*, una categoría que incluye al Perceptron de Rosenblatt. Otros tipos de redes neuronales no tenían esta limitación. De todas maneras, el libro logró acabar con la mayor parte de la financiación para la investigación sobre redes neuronales en la década de 1970. Este campo volvió a resurgir en la década de 1980 mediante intentos por utilizar modelos de neuronas biológicas que supuestamente eran más realistas, así como modelos que evitaban las limitaciones derivadas del teorema del Perceptron de Minsky-Papert. No obstante, la capacidad del neocórtex para resolver el problema de la invarianza, uno de sus puntos fuertes fundamentales, era algo que siguió estando vetado para el renacido campo del conexionismo.

## *Codificación dispersa: cuantificación vectorial*

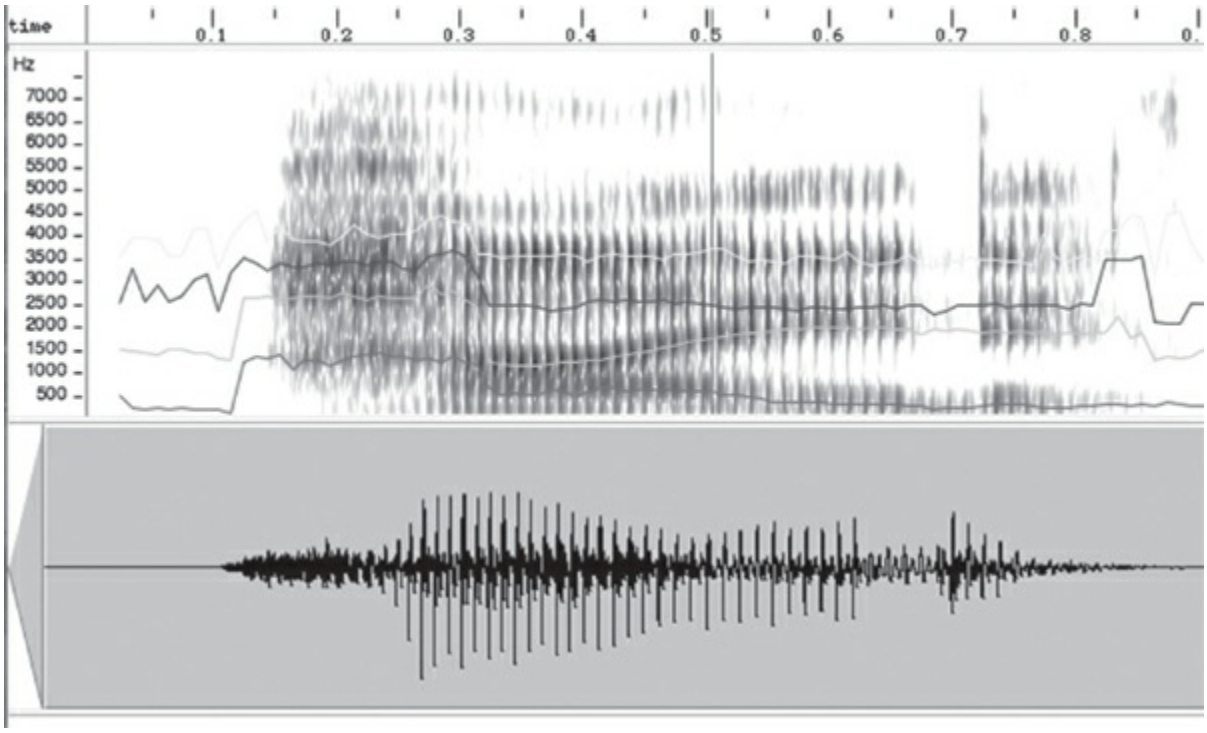
A principios de la década de 1980 inauguré un proyecto dedicado a otro problema clásico del reconocimiento de patrones: la comprensión del habla humana. Al principio usamos las estrategias de la inteligencia artificial (IA) tradicional. Así, programábamos directamente conocimientos expertos sobre las unidades fundamentales del habla (los fonemas) y reglas que ideaban los lingüistas sobre cómo las personas concatenan los fonemas para formar palabras y frases. Cada fonema posee una frecuencia propia de patrones. Por ejemplo, sabíamos que sonidos vocales tales como “e” y “a” se caracterizan por ciertas frecuencias resonantes llamadas formantes y que cada fonema posee una proporción de formantes característica. Por su parte, los sonidos líquidos como el de la “z” y la “s” se caracterizan por hacer brotar sonidos que abarcan muchas frecuencias.

Registrábamos el habla en forma de onda y luego la convertíamos en múltiples bandas de frecuencia (percibidas como tonos) usando un banco de filtros de frecuencia. El resultado de esta transformación podía ser visualizado y recibía el nombre de espectrograma (véase la página siguiente).

El banco de filtros copaba la función realizada por la cóclea, que representa el paso inicial de nuestro procesamiento biológico del sonido. Primero, el software identificaba fonemas distinguiendo entre patrones de frecuencias, y luego identificaba palabras basándose en identificaciones de secuencias de fonemas características.




**Un espectrograma de tres vocales. De izquierda a derecha: [i] como en "appreciate", [u] como en "aco*u*stic" y [a] como en "ah"*. El eje Y representa la frecuencia del sonido. Cuanto más oscura la banda, más energía acústica existe en dicha frecuencia.**




**Un espectrograma de una persona diciendo la palabra "hide"*. Las líneas horizontales muestran los formantes, que son frecuencias sostenidas que poseen una energía especialmente alta (10).**

El resultado fue parcialmente exitoso. Pudimos entrenar nuestro dispositivo para que aprendiera los patrones correspondientes a una persona que usase un vocabulario de tamaño moderado compuesto de miles de palabras. Cuando intentamos que reconociera decenas de miles de palabras, se encargara de varios hablantes y permitiera un habla completamente continua (es decir, el habla sin pausas entre las palabras), nos dimos con el problema de la invarianza. Diferentes personas enunciaban el mismo fonema de forma diferente. Por ejemplo, en una persona el fonema "e" puede sonar como el fonema "a" de otra. Incluso la misma persona era

inconsistente en su manera de pronunciar un fonema en particular. A menudo, el patrón de un fonema se veía afectado por otros fonemas próximos, y muchos fonemas eran completamente excluidos. La pronunciación de las palabras (es decir, la forma en la que los fonemas se concatenan para formar las palabras) también era altamente variable y dependiente del contexto. Las reglas lingüísticas que habíamos programado se vinieron abajo y no pudimos hacer frente a la extremada variabilidad del lenguaje hablado.

En aquel momento tuve claro que la esencia del reconocimiento humano de patrones y de conceptos se basaba en jerarquías. Esto es muy obvio en el lenguaje humano, que conforma una complicada jerarquía de estructuras. ¿Pero cuál es el elemento que se encuentra en la base de las estructuras? Esta fue la primera pregunta que me planteé cuando me puse a investigar sobre las maneras para reconocer automáticamente el habla humana normal y corriente.

El sonido entra por el oído en forma de vibración del aire y las aproximadamente 3 000 células ciliadas internas en el interior de la cóclea las convierten en múltiples bandas de frecuencia. Todas las células ciliadas están sintonizadas en una frecuencia en particular (téngase en cuenta que percibimos las frecuencias como tonos) y todas actúan como filtros de frecuencias que emiten una señal cuando quiera que un sonido se produzca en su frecuencia de resonancia o en una frecuencia cercana. Por tanto, al abandonar la cóclea humana, el sonido viene representado por aproximadamente 3 000 señales individuales, cada una de las cuales representa la intensidad de tiempo variable correspondiente a una estrecha banda de frecuencias entre las que se da un importante nivel de solapamiento.

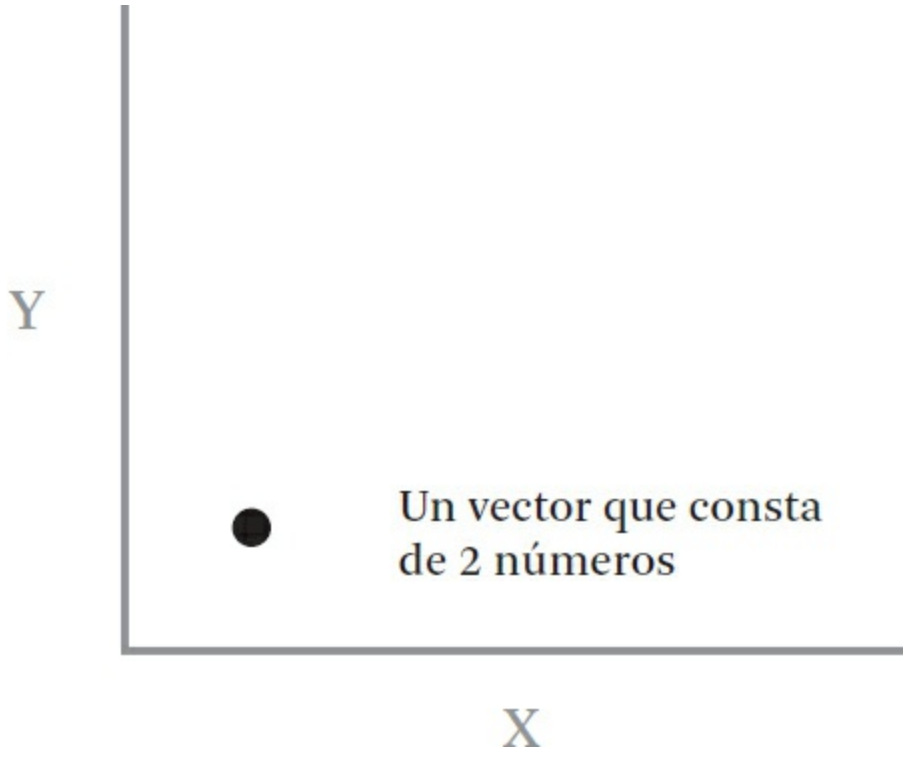
Aunque era evidente que el cerebro funcionaba de forma masivamente paralela, me parecía imposible que realizara un emparejamiento de patrones que constara de 3 000 señales auditivas diferentes. Dudaba sobre la posibilidad de que la evolución hubiera sido tan ineficiente. Ahora sabemos que antes de que las señales del sonido alcancen el neocórtex, tiene lugar una importante reducción de datos en el nervio auditivo.

En el caso de nuestros reconocedores del habla mediante software, también usamos filtros que hacían las veces de software (en concreto 16 filtros que luego aumentamos hasta 32, ya que descubrimos que superar esta cifra no reportaba ningún beneficio considerable). Así, en nuestro sistema cada punto temporal venía representado por 16 números. Teníamos que reducir estos 16 chorros de datos a uno a la vez que enfatizábamos las características importantes para el reconocimiento del habla.

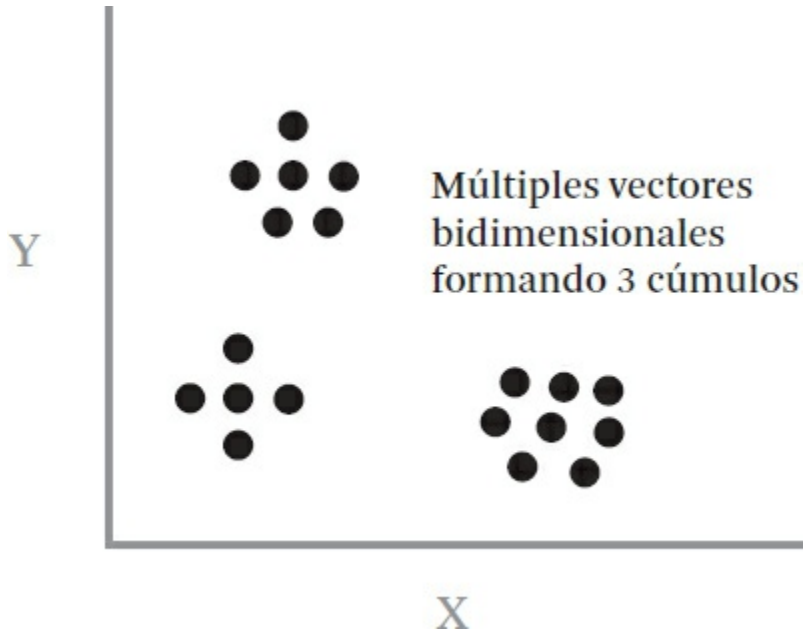
Para conseguir esto utilizamos una técnica matemática óptima llamada cuantificación vectorial. Téngase en cuenta que en un determinado punto temporal el sonido de por lo menos un oído venía representado por nuestro software mediante 16 números diferentes, es decir, por el output de 16 filtros de frecuencia. (En el sistema

auditivo humano esta cifra sería de 3 000, correspondiente al output de las 3 000 células ciliadas internas). En términos matemáticos, cada conjunto de números de estas características (3 000 en el caso biológico o 16 en nuestra implementación por software) recibe el nombre de vector.

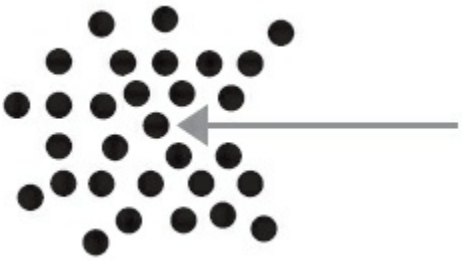
En aras de la simplicidad, consideremos el proceso de cuantificación vectorial mediante vectores de dos números. Cada vector puede ser considerado un punto en un espacio bidimensional.




Si contamos con una muestra muy amplia de dichos vectores y los representamos, es posible que notemos cómo se forman cúmulos.



Para identificar los cúmulos tenemos que decidir cuántos vamos a tolerar. En nuestro proyecto solíamos permitir 1 024 cúmulos, de manera que podíamos numerarlos y asignar a cada cúmulo una etiqueta de 10 bits, ya que $2^{10}$ = 1 024. Nuestra muestra de vectores representa la diversidad que esperamos. A modo de tentativa, asignamos los primeros 1 024 vectores a cúmulos de un solo punto. Entonces tomamos en consideración el vector número 1 025 e identificamos su punto más cercano. Si dicha distancia es mayor que la distancia más pequeña entre cualquier par de los 1 024 puntos, lo consideramos como el principio de un nuevo cúmulo. Luego colapsamos los dos cúmulos de un solo punto más cercanos entre ellos en un cúmulo individual. De esta manera seguimos teniendo 1 024 cúmulos. Después de procesar el vector número 1 025, uno de dichos cúmulos posee ahora más de un solo punto. Continuamos procesando puntos de esta manera, siempre manteniendo los 1 024 cúmulos. Después de procesar todos los puntos, representamos cada uno de los cúmulos de muchos puntos a partir del centro geométrico de los puntos en dicho cúmulo.



Un cúmulo de puntos. Lo representamos usando un solo punto que es el centro geométrico de todos los puntos del cúmulo.

Continuamos con este proceso iterativo hasta abarcar todos los puntos de la muestra. Normalmente procesábamos millones de puntos en el interior de 1 024 ($2^{10}$) cúmulos; también hubo veces en que usamos 2 048 ($2^{11}$) o 4 096 ($2^{12}$) cúmulos. Cada cúmulo viene representado por un vector que se encuentra en el centro geométrico de todos los puntos de dicho cúmulo. Así, el total de las distancias desde cada uno de los puntos del cúmulo hasta el punto central del cúmulo es todo lo pequeño posible.

El resultado de esta técnica es que en vez de tener los millones de puntos con los que comenzamos (y un número todavía mayor de posibles puntos), los datos se reducen hasta solo 1 024 puntos que utilizan el espacio de posibilidades de forma óptima. Las partes del espacio que no son usadas nunca no son asignadas a ningún cúmulo.

Entonces asignamos un número a cada cúmulo (en nuestro caso, de 0 a 1 023). Ese número es la representación reducida, “cuantificada”, de dicho cúmulo, razón por la cual la técnica recibe el nombre de cuantificación. Cualquier vector de input nuevo que llegue en el futuro se representará mediante el número del cúmulo cuyo punto central sea el más cercano a este nuevo vector de input.

En este momento se puede precomputar una tabla con la distancia desde el punto

central de cada uno de los cúmulos hasta todos los demás puntos centrales. Por tanto, obtenemos instantáneamente la distancia de este nuevo vector de input (cosa que representamos mediante este punto cuantificado; en otras palabras, mediante el número de cúmulos a los que este nuevo punto está más cercano) hasta todos los demás cúmulos. Dado que estamos representando solamente puntos con respecto a su cúmulo más cercano, conoceremos la distancia desde este punto hasta cualquier otro punto posible que pudiera aparecer.

La técnica de más arriba la he descrito usando vectores con solo dos números cada uno, pero trabajar con vectores de 16 elementos es completamente análogo al ejemplo más sencillo. Dado que elegimos vectores con 16 números que representaban 16 frecuencias de banda diferentes, cada punto de nuestro sistema era un punto en un espacio de 16 dimensiones. Para nosotros es muy difícil imaginar un espacio con más de tres dimensiones (quizás cuatro, si incluimos el tiempo), pero las matemáticas no sufren dichas inhibiciones.

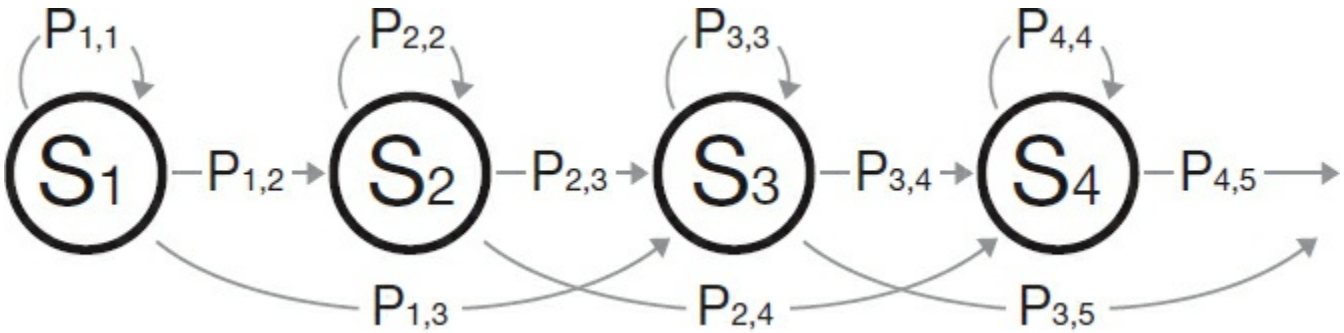
Con este proceso hemos logrado cuatro cosas. Primero, hemos reducido enormemente la complejidad de los datos. Segundo, hemos reducido datos de 16 dimensiones a datos unidimensionales (es decir, ahora cada muestra es un solo número). Tercero, hemos mejorado nuestra capacidad para encontrar características invariables ya que hemos enfatizado las porciones del espacio de posibles sonidos que verbalizan la mayor parte de la información. La mayor parte de las combinaciones de frecuencias son físicamente imposibles o por lo menos muy improbables, de manera que no hay razón para dar el mismo espacio a las combinaciones improbables de inputs que a las que son probables. Esta técnica reduce los datos a posibilidades igualmente probables. El cuarto beneficio es que podemos usar reconocedores de patrones unidimensionales aunque los datos originales consten de muchas más dimensiones. A su vez, esto demostró ser la estrategia más eficiente para aprovechar los recursos informáticos disponibles.

## *Leer la mente mediante los modelos ocultos de Márkov*

Con la cuantificación vectorial simplificamos los datos para que enfatizaran características fundamentales. Sin embargo, seguíamos necesitando una manera de representar la jerarquía de características invariables que dotaran de coherencia a la nueva información. En aquel momento (principios de la década de 1980), después de haber trabajado en el campo del reconocimiento de patrones durante veinte años, yo ya sabía que las representaciones unidimensionales eran mucho más potentes, eficientes y apropiadas para los resultados invariables. A principios de la década de 1980 no se sabía mucho del neocórtex, pero basándome en mi propia experiencia

con una serie de problemas relacionados con el reconocimiento de patrones, supuse que también era probable que el cerebro redujera sus datos multidimensionales, bien a partir de los ojos, bien a partir de los oídos o de la piel, usando una representación unidimensional (sobre todo a medida que los conceptos se encuentran más arriba en la jerarquía del neocórtex).

En el caso del problema del reconocimiento del habla, la organización de la información en las señales del habla parecía seguir una jerarquía de patrones en la que cada patrón venía representado por una sucesión de elementos que avanzaban hacia adelante. Todos los elementos de un patrón podrían ser un patrón de nivel más bajo o una unidad fundamental de input (que en el caso del reconocimiento del habla haría las veces de vectores cuantificados).




**Un ejemplo sencillo de una capa de un modelo oculto de Márkov. Entre $S_1$ y $S_4$ vienen representados los estados internos "ocultos". Cada transición $P_{i,j}$ representa la probabilidad de pasar desde un estado $S_i$ hasta un estado $S_j$. Dichas probabilidades vienen determinadas por el aprendizaje que realiza el sistema a partir de los datos con los que ha sido entrenado (que incluyen el uso real). A estas probabilidades se les atribuye una nueva secuencia (como por ejemplo una nueva expresión hablada) con el objeto de determinar la probabilidad de que este modelo haya producido dicha secuencia.**

Reconocerá que esta situación es coherente con el modelo del neocórtex que he expuesto anteriormente. El habla humana, por tanto, se produce en el cerebro mediante una jerarquía de patrones lineales. Si simplemente se pudiera examinar estos patrones en el interior del cerebro de la persona que está hablando, sería fácil comparar su habla con los patrones cerebrales y comprender lo que esa persona estuviera diciendo. Por desgracia no tenemos acceso directo al cerebro de una persona que hable, la única información que tenemos es lo que está diciendo. Por supuesto, esa es la razón de ser del lenguaje hablado: el orador comparte un trozo de su mente mediante el habla.

De manera que me pregunté: ¿hay alguna técnica matemática que nos permita, basándonos en sus palabras, inferir los patrones en el cerebro del orador? Obviamente, una expresión no sería suficiente, pero si tuviéramos una gran cantidad de muestras, ¿podríamos usar esa información para básicamente leer los patrones en el interior del neocórtex del orador (o por lo menos para formular algo matemáticamente equivalente que nos permitiera reconocer nuevas expresiones?)

A menudo la gente no es capaz de apreciar lo poderosas que pueden ser las matemáticas. Tenga en cuenta que nuestra capacidad para buscar gran parte del conocimiento humano en una fracción de segundo mediante buscadores se basa en una técnica matemática. En el caso del problema sobre el reconocimiento de patrones al que me enfrenté a comienzos de la década de 1980, resultó que la técnica de los modelos ocultos de Márkov se adecuaba perfectamente. El matemático ruso Andrei Andreyevich Márkov (1856–1922) construyó una teoría matemática consistente en secuencias jerárquicas de estados. Este modelo se basaba en la posibilidad de cruzar los estados de una cadena, y en el caso de conseguirlo, hacer que se disparara un estado perteneciente al siguiente nivel más alto de la jerarquía. ¿Resulta familiar?

El modelo de Márkov incluía las probabilidades de que se dieran cada uno de los estados. Además, propuso la hipótesis de una situación en la que un sistema poseyera una jerarquía de secuencias lineales de estados de estas características, con excepción de aquellos estados que fuera imposible examinar directamente, de ahí el nombre de modelos *ocultos* de Márkov. El nivel más bajo de la jerarquía emite señales que son lo único que podemos ver. Márkov proporciona una técnica matemática para calcular, basándose en el output observado, cuáles deben ser las probabilidades de cada transición. El método fue posteriormente perfeccionado por Norbert Wiener en 1923. El perfeccionamiento llevado a cabo por Wiener también proporcionó una manera de determinar las conexiones en el modelo de Márkov, donde cualquier conexión con una probabilidad demasiado baja era considerada como no existente. Básicamente, así es cómo el neocórtex humano recorta conexiones: si nunca o muy pocas veces son utilizadas, se las considera improbables y son podadas. En nuestro caso, el output observado es la señal hablada creada por la persona al hablar, y las probabilidades de estado y las conexiones del modelo de Márkov constituyen la jerarquía neocortical que dio lugar al propio output.

Yo concebí un sistema en el cual se tomarían muestras del habla humana, se aplicaría la técnica del modelo oculto de Márkov para inferir una jerarquía de estados dotada de conexiones y probabilidades (esencialmente un neocórtex simulado para producir habla) y luego se usaría este mecanismo de estados jerárquicos inferidos para reconocer nuevas expresiones. Para crear un sistema independiente del hablante se usarían muestras procedentes de muchos individuos diferentes que entrenaran a los modelos ocultos de Márkov. Al añadir el elemento de las jerarquías para representar la naturaleza jerárquica de la información contenida en el lenguaje, se conseguían verdaderos modelos ocultos de Márkov (HHMMs)*.

Mis colegas de Kurzweil Applied Intelligence se mostraron escépticos sobre la probabilidad de que esta técnica funcionase, ya que se trataba de un método autoorganizativo que evocaba las redes neuronales que habían caído en desgracia y

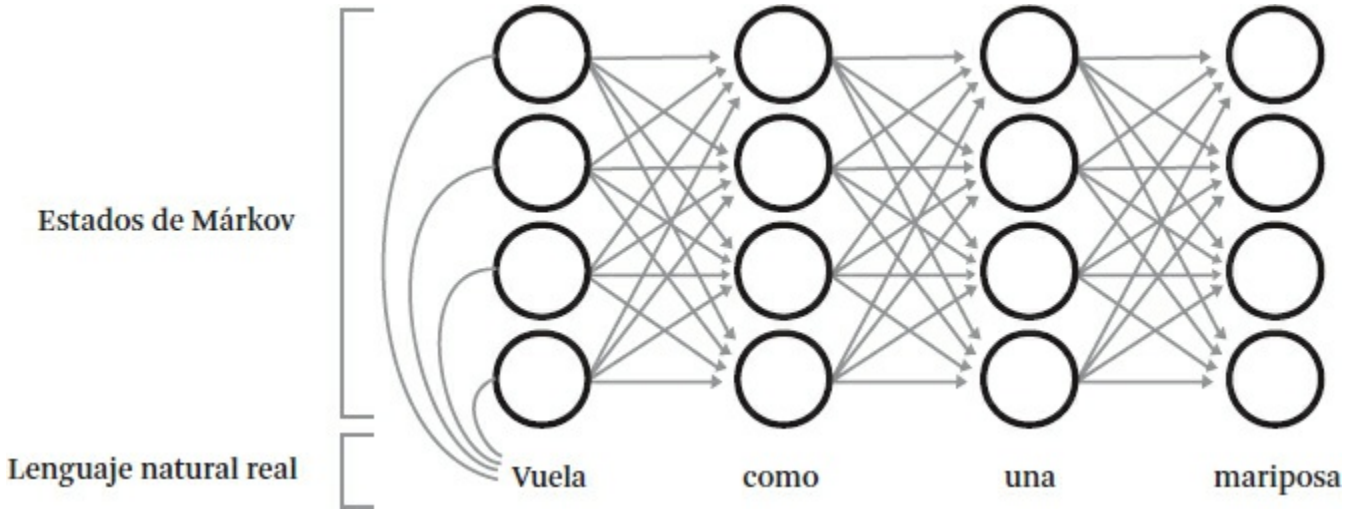
con las que habíamos tenido poco éxito. Yo señalé que el mecanismo de un sistema de red neuronal es fijo y no se adapta al input (el peso sí que se adapta, pero las conexiones no). En el sistema del modelo de Márkov, siempre que estuviera correctamente montado, el sistema podaría las conexiones no utilizadas para adaptarse a la topología.

Establecí lo que fue considerado como un "programa de desarrollo avanzado"* (un término organizativo que se aplica a un proyecto que se sale de los senderos trillados y que cuanta con pocos recursos formales). Constaba de mí mismo, de un programador a tiempo parcial y de un ingeniero electrónico que tenía que crear el banco de filtros de frecuencia. Para sorpresa de mis colegas, nuestro esfuerzo resultó ser muy fructífero y logramos reconocer con mucha precisión habla que contenía un vocabulario muy amplio.

Después de este experimento, todos nuestros esfuerzos posteriores en el campo del reconocimiento del habla se han basado en los modelos ocultos jerárquicos de Márkov. De forma aparentemente independiente, otras compañías de reconocimiento del habla descubrieron la valía de este método, y desde mediados de la década de 1980 la mayor parte del trabajo sobre el reconocimiento automático del habla se ha basado en esta estrategia. Los modelos ocultos de Márkov son asimismo utilizados en la síntesis del habla, ya que hay que tener en cuenta que nuestra jerarquía cortical biológica no solo se usa para reconocer input, también se utiliza para producir output, por ejemplo, en el caso del habla y del movimiento físico.

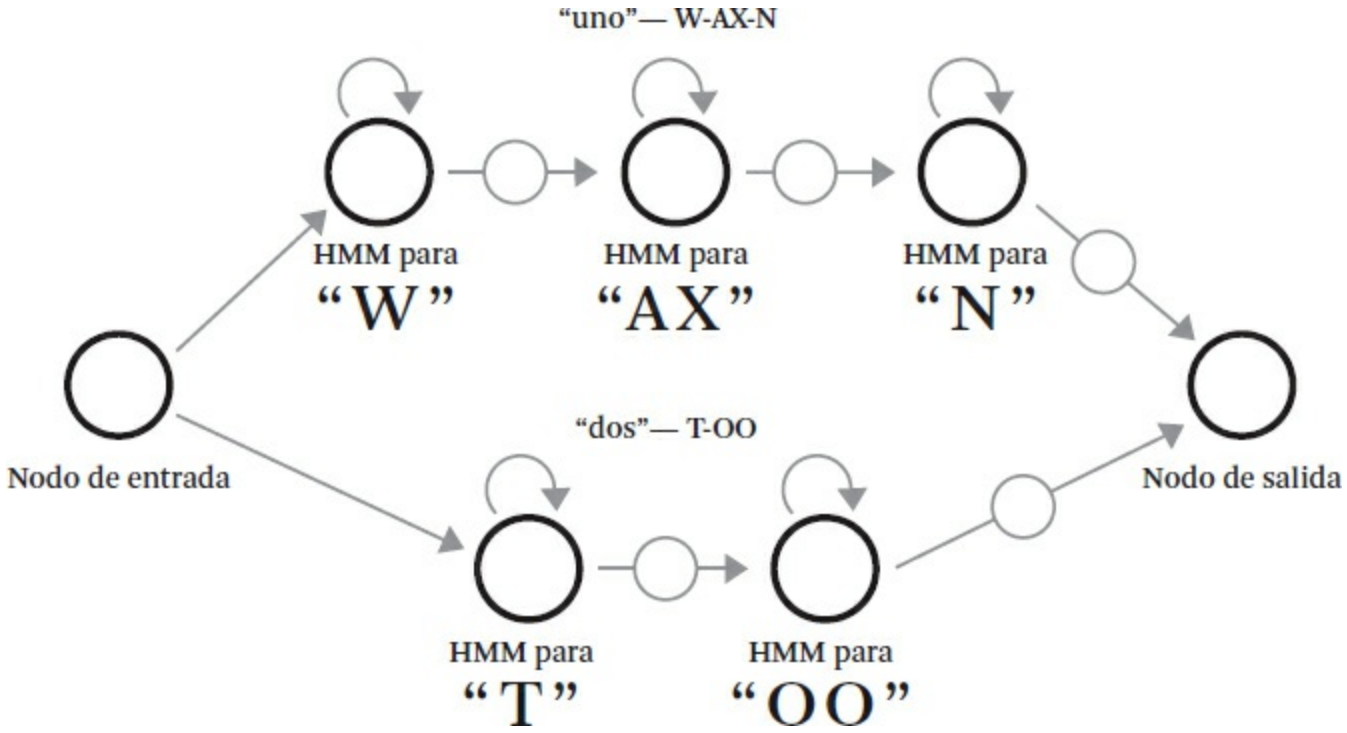
HHMMs también son utilizados en sistemas que comprenden el significado de las frases en lenguaje natural, lo cual representa un ascenso en la jerarquía conceptual.

Para comprender cómo funciona el método HHMM, empecemos por un mecanismo consistente en todos los estados de transición posibles. En esto, el método de cuantificación vectorial descrito anteriormente es fundamental, ya que de otra manera habría que considerar demasiadas posibilidades.



**Los estados ocultos de Márkov y las posibles transiciones para producir una secuencia de palabras en texto en lenguaje natural.**

He aquí una posible topología inicial simplificada:




**Una topología de un modelo oculto de Márkov sencillo para reconocer dos palabras habladas*.**

Las expresiones que sirven de muestra son procesadas una por una. En cada caso modificamos iterativamente las probabilidades de las transiciones para que se refleje mejor el input de muestra que acabamos de procesar. Los modelos de Márkov usados en el reconocimiento del habla codifican la probabilidad de que patrones específicos de sonido sean encontrados en cada fonema, cómo los fonemas se influencian los unos a los otros y el orden probable de los fonemas. El sistema también puede incluir mecanismos de probabilidad en los niveles superiores de la estructura del lenguaje, como en el caso del orden de las palabras, la inclusión de frases hechas y así sucesivamente a lo largo de la jerarquía del lenguaje.

Mientras que los sistemas de reconocimiento del habla anteriores incorporaban reglas específicas sobre las estructuras de fonemas y secuencias expresamente codificadas por lingüistas humanos, al nuevo sistema basado en HHMM no se le dijo expresamente que en inglés existen 44 fonemas, ni las secuencias de vectores que son más probables para cada fonema, ni qué secuencias de fonemas eran más probables que otras. Dejamos que el sistema descubriera estas "reglas" por sí mismo a partir de miles de horas de transcripción de datos del habla humana. La ventaja de esta estrategia sobre las reglas codificadas a mano es que los modelos desarrollan reglas probabilísticas de las que a menudo los expertos humanos no están al tanto. Así, nos dimos cuenta de que muchas de estas reglas aprendidas automáticamente por el

sistema a partir de los datos diferían de forma sutil pero importante de las reglas establecidas por los expertos humanos.

Una vez que el mecanismo estuvo entrenado, empezamos a intentar reconocer el habla teniendo en cuenta los caminos alternativos a través del mecanismo y eligiendo el camino más probable según la secuencia de vectores de input que hubiéramos detectado. En otras palabras, si percibíamos una secuencia de estados que probablemente hubiera producido la expresión en cuestión, llegábamos a la conclusión de que dicha expresión provenía de esa secuencia cortical. Este neocórtex simulado basado en HHMM incluía palabras etiqueta, de manera que era capaz de sugerir una transcripción de lo que hubiera oído.

A continuación estuvimos en disposición de mejorar aún más los resultados mediante la continuación del entrenamiento del mecanismo al mismo tiempo que lo usábamos en labores de reconocimiento. Tal y como he comentado, el reconocimiento y el aprendizaje simultáneos también se producen en todos los niveles de la jerarquía neocortical biológica.

## *Algoritmos evolutivos (genéticos)*

Queda por hacer otra consideración importante: ¿Cómo configurar los muchos parámetros que controlan el funcionamiento de un sistema de reconocimiento de patrones? Dichos parámetros podrían incluir el número de vectores que vayamos a permitir en la fase de cuantificación vectorial, la topología inicial de los estados jerárquicos anterior a la poda realizada durante la fase de entrenamiento del modelo oculto de Márkov, el umbral de reconocimiento de cada nivel de la jerarquía, los parámetros que controlan el manejo de los parámetros de tamaño y muchos otros. Podemos determinarlos basándonos en nuestra intuición, pero los resultados estarían muy lejos de ser óptimos.

A estos parámetros los llamamos “los parámetros de dios”, ya que son configurados antes que el método autoorganizativo que determina la topología de los modelos ocultos de Márkov (o, en el caso biológico, antes de que la persona aprenda las lecciones creando conexiones similares en su jerarquía cortical). Quizá no se trate de un nombre apropiado, ya que estos detalles del diseño original basados en el ADN vienen determinados por la evolución biológica, aunque algunos puedan ver la mano de dios en dicho proceso (y aunque considero que la evolución es un proceso espiritual, esta es una discusión que perteneces al capítulo 9).

Al configurar estos “parámetros de dios” en nuestro sistema jerárquico simulado de reconocimiento y aprendizaje, volvimos a tomar como ejemplo la naturaleza y decidimos someterlos a evolución (en nuestro caso, mediante una evolución

simulada). Para ello, utilizamos los llamados algoritmos genéticos o evolutivos (AGs), que incluyen mutaciones y reproducción sexual simuladas.

He aquí una descripción simplificada de cómo funciona este método. Primero, determinamos una manera de codificar las posibles soluciones de un problema en concreto. Si el problema es la optimización de los parámetros de diseño de un circuito, hacemos una lista con todos los parámetros que caracterizan el circuito según el número específico de bits asignado a cada parámetro. Esta lista se considera el código genético del algoritmo genético. Después, generamos aleatoriamente miles o más códigos genéticos. Cada uno de estos códigos genéticos, que representan un conjunto de los parámetros de diseño, es considerado como un organismo "solución" simulado.

Entonces evaluamos cada organismo simulado en un ambiente simulado usando un método en concreto para evaluar cada conjunto de parámetros. Esta evaluación es clave para el éxito de un algoritmo genético. En nuestro ejemplo, ejecutaríamos todos los programas generados por estos parámetros y los juzgaríamos según los criterios apropiados (si completó o no la tarea, cuánto tardó, etc). A los organismos que representen la mejor solución (los mejores diseños) se les permite sobrevivir y el resto es eliminado.

En este punto hacemos que todos los supervivientes se multipliquen a sí mismos hasta que alcancen el mismo número de criaturas solución. Esto se realiza simulando la reproducción sexual. En otras palabras, creamos nuevos descendientes en el punto en el que cada nueva criatura desarrolla una parte de su código genético a partir de un progenitor y otra parte a partir del segundo progenitor. Por lo general, no se distingue entre organismos macho o hembra, basta con generar un descendiente a partir de dos progenitores cualquiera, de manera que en este caso se trata básicamente de matrimonios del mismo sexo. Puede que no sea tan interesante como la reproducción sexual del mundo natural, pero lo relevante es que hay dos progenitores. A medida que estos organismos simulados se multiplican, se permite cierto nivel de mutación (cambio aleatorio) en los cromosomas.

Ahora ya tenemos una generación producto de la evolución simulada, y repetimos estos pasos para cada generación subsecuente. Al final de cada generación se determina la mejora sufrida por los diseños, es decir, se calcula la mejora media en la función de evaluación correspondiente a todos los organismos supervivientes. Cuando el grado de mejora en la evaluación de las criaturas de diseño se vuelve muy pequeña de generación a generación, se detiene este ciclo iterativo y se usan el/los mejor(es) diseño(s) de la última generación. (Para una descripción algorítmica de los algoritmos genéticos, véase la nota (11)).

La clave de un algoritmo genético es que los diseñadores humanos no programan una solución directamente. En vez de eso, se permite que surja una a través de un proceso iterativo de competición simulada y de mejora. La evolución biológica es

lista pero lenta, de manera que para aumentar su inteligencia aceleramos en gran medida sus pesados andares. El ordenador es lo suficientemente rápido como para simular muchas generaciones en cuestión de horas o días, y ocasionalmente han de estar en funcionamiento durante algunas semanas para simular cientos de miles de generaciones. Sin embargo, este proceso iterativo solo tenemos que recorrerlo una vez. En cuanto hayamos permitido que esta evolución simulada siga su curso, podemos aplicar rápidamente las reglas fruto de la evolución y altamente perfeccionadas en problemas reales. En el caso de nuestros sistemas de reconocimiento del habla, las utilizamos para evolucionar la topología inicial del mecanismo y otros parámetros fundamentales. Por lo tanto, utilizamos dos métodos autoorganizativos: un algoritmo genético (AG) para simular la evolución biológica que dio lugar a un diseño cortical concreto y HHMMs para simular la organización cortical que acompaña al aprendizaje humano.

Otro importante requisito para el éxito de un AG es un método válido para evaluar toda solución posible. Esta evaluación tiene que ser hecha rápidamente, ya que tiene que tener en cuenta muchos miles de soluciones posibles por cada generación de la evolución simulada. Los AGs son expertos en abordar los problemas con demasiadas variables como para calcular de forma precisa las soluciones analíticas. Por ejemplo, el diseño de un motor puede requerir de más de cien variables y tiene que satisfacer docenas de restricciones. Los AGs utilizados por los investigadores de General Electric han sido capaces de crear diseños de motores para aviones que cumplían con las restricciones de una forma más precisa a como lo hacen los métodos convencionales.

Sin embargo, al usar AGs hay que tener cuidado con lo que se desea. Un algoritmo genético fue utilizado para resolver un problema de apilamiento de bloques y creó una solución perfecta… si exceptuamos que dicha solución constaba de miles de pasos. Los programadores humanos olvidaron incluir en su función de evaluación una minimización en el número de pasos.

El proyecto Electric Sheep* de Scott Drave es un AG que produce arte. La función de evaluación utiliza evaluadores humanos en una colaboración de código abierto en la que participan muchos miles de personas. El arte se mueve a través del tiempo y esto se puede ver en electricsheep.org.

En el caso del reconocimiento del habla, la combinación de los algoritmos genéticos y los modelos ocultos de Márkov ha funcionado extremadamente bien. La simulación de la evolución mediante un AG permitió mejorar sustancialmente el rendimiento de los mecanismos basados en HHMM. Lo que la evolución produjo fue muy superior al diseño original basado en nuestra intuición.

Entonces experimentamos con la introducción de sucesiones de pequeñas variaciones en el sistema general. Por ejemplo, llevábamos a cabo perturbaciones

(pequeños cambios aleatorios) en el input. Otro cambio consistía en hacer que modelos de Márkov colindantes "se filtraran" el uno al otro, cosa que hacía que los resultados de un modelo de Márkov influenciaran los modelos "cercanos". Aunque en aquel momento no nos dimos cuenta, el tipo de ajustes con los que experimentamos son muy similares a los tipos de modificaciones que tienen lugar en las estructuras corticales biológicas.

Al principio, dichos cambios dañan el rendimiento (si medimos este mediante la precisión del reconocimiento). Pero si volvemos a ejecutar la evolución (es decir, si volvemos a ejecutar el AG) con estas alteraciones, el AG hará que el sistema se adapte consecuentemente y lo optimizará según las modificaciones introducidas. Por lo general, esto hace que el rendimiento se recupere. Si luego eliminamos los cambios introducidos, el rendimiento vuelve a degradarse, ya que el sistema ha evolucionado para compensar los cambios. El sistema adaptado se vuelve dependiente de los cambios.

Un tipo de alteración que de hecho mejoraba el rendimiento después de volver a ejecutar el AG era la introducción de pequeños cambios aleatorios en el input. La razón para ello es el bien conocido problema del "sobreajuste"* en los sistemas autoorganizativos. Existe el peligro de que un sistema así generalice en exceso los ejemplos específicos contenidos en la muestra utilizada para el entrenamiento. Cuantos más ajustes aleatorios se realicen en el input, más patrones invariantes sobreviven en los datos y por tanto el sistema aprende estos patrones más profundos. Esto era de ayuda solo si volvíamos a ejecutar el AG con la función de aleatoriedad activada.

Esto provoca un dilema en nuestra comprensión de los circuitos corticales biológicos. Por ejemplo, se ha comprobado que de hecho puede existir una pequeña cantidad de filtración desde una conexión cortical a otra producto de la manera en que se forman las conexiones biológicas, ya que aparentemente la electroquímica de los axones y de las dendritas está sujeta a los efectos electromagnéticos de las conexiones cercanas. Supongamos que somos capaces de ejecutar un experimento en el que elimináramos este efecto en un cerebro real. Se trata de algo difícil de conseguir, pero no de algo imposible. Supongamos que realizamos dicho experimento y descubrimos que los circuitos corticales funcionan con menor eficiencia sin esta filtración neuronal. Entonces podríamos concluir que este fenómeno es fruto de un diseño muy inteligente de la evolución y que fue fundamental para que el córtex alcanzara su nivel de rendimiento. Además, dada la intrincada influencia que se da entre las conexiones, podríamos señalar que dicho resultado demuestra que el modelo del ordenamiento de los patrones que fluyen hacia arriba por la jerarquía conceptual y de las predicciones que fluyen hacia abajo por dicha jerarquía era en realidad mucho más complicado de lo que podía parecer.

Sin embargo, esa no sería una conclusión necesariamente precisa. Consideremos

nuestra experiencia con un córtex simulado basado en HHMMs en el que implementáramos una modificación que fuera muy similar a la comunicación cruzada interneuronal. Si entonces ejecutáramos la evolución con este fenómeno incluido, el rendimiento se recuperaría, ya que el proceso evolutivo se adaptaría a él. Si entonces eliminámos la comunicación cruzada, el rendimiento volvería a resentirse. En el caso biológico, la evolución (es decir, la evolución biológica) fue "ejecutada" incluyendo este fenómeno. Por tanto, los parámetros detallados del sistema han sido configurados por la evolución biológica para que dependieran de estos factores, de manera que cambiarlos afectaría al rendimiento negativamente a no ser que volviéramos a ejecutar la evolución. Esto es posible en el mundo de la simulación, donde la evolución solo tarda días o semanas, pero en el mundo biológico se requerirían decenas de miles de años.

Entonces, ¿cómo podemos saber si una característica en particular del diseño del neocórtex biológico es una innovación vital introducida por la evolución biológica, es decir, que es determinante para nuestro nivel de inteligencia, o meramente un artefacto del que el diseño del sistema es dependiente pero sin el que también hubiera podido evolucionar? Podemos dar respuesta a esta cuestión simplemente ejecutando una evolución simulada con y sin estas variaciones concretas en los detalles del diseño, por ejemplo, con y sin comunicación cruzada. Incluso podemos hacerlo con la evolución biológica si examináramos la evolución de una colonia de microorganismos en la que las generaciones se suceden en horas; sin embargo, esto no resulta práctico en organismos complejos como los humanos. Este es otro de los muchos inconvenientes de la biología.

Volviendo a nuestro trabajo en el campo del reconocimiento del habla, descubrimos que si ejecutábamos la evolución (es decir, un AG) *por separado* sobre el diseño inicial de (1) los modelos ocultos jerárquicos de Márkov que modelizaban la estructura interna de los fonemas y (2) los HHMMs que modelizaban las estructuras de las palabras y frases, obteníamos resultados que eran todavía mejores. Ambos niveles del sistema usaban HHMMs, pero el AG desarrollaba variaciones en el diseño entre dichos niveles. Además, esta estrategia seguía permitiendo la modelización de los fonemas que tiene lugar entre los dos niveles, como por ejemplo el arrastre de los fonemas que suele ocurrir cuando unimos ciertas palabras (por ejemplo, "¿cómo estás?" puede convertirse en "¿comoestás?").

Es probable que un fenómeno similar tenga lugar en diferentes regiones corticales biológicas, ya que han desarrollado pequeñas diferencias basadas en los tipos de patrones a los que tienen que enfrentarse. Aunque todas estas regiones usan el mismo algoritmo neocortical esencial, la evolución biológica ha tenido tiempo suficiente para afinar el diseño de cada uno de ellos y que así resultaran óptimos para sus patrones particulares. Sin embargo, tal y como ya he comentado, los

neurocientíficos y los neurólogos han detectado una importante plasticidad en estas áreas, lo cual respalda la idea de un algoritmo neocortical general. Si en cada región, los métodos fundamentales fueran radicalmente diferentes, entonces esta intercambiabilidad entre regiones corticales no sería posible.

Los sistemas que creamos durante nuestra investigación utilizando esta combinación de métodos autoorganizativos tuvieron mucho éxito. En el campo del reconocimiento del habla, pudieron por primera vez soportar habla completamente continua y vocabularios relativamente ilimitados. Fuimos capaces de alcanzar un grado de precisión elevado en una amplia variedad de hablantes, acentos y dialectos. El actual estado de cosas mientras se escribe este libro viene representado por un producto llamado Dragon Naturally Speaking (versión 11.5) para PC perteneciente a Nuance (que anteriormente fue Kurzweil Computer Products). Si hay personas escépticas sobre el rendimiento de los sistemas de reconocimiento del habla actuales, recomiendo que lo prueben. Los niveles de precisión a menudo son del 99% o superiores después de unos minutos de entrenamiento con su voz realizando habla continua e incorporando vocabularios relativamente ilimitados. Dragon Dictation es un "free app" más sencillo pero igualmente asombroso para iPhone que no necesita de entrenamiento con voz. Siri, el asistente personal de los actuales iPhones de Apple, utiliza la misma tecnología de reconocimiento del habla dotada de avances que le permiten comprender el lenguaje natural.

El rendimiento de estos sistemas da testimonio del poder de las matemáticas. Pese a no tener acceso directo al cerebro de las personas, mediante ellas computamos lo que está pasando en el neocórtex de un hablante, lo cual es un paso fundamental en el reconocimiento de lo que dicho hablante está diciendo y, en el caso de sistemas como Siri, de lo que significan sus expresiones. Podríamos preguntarnos, ¿si de hecho pudiéramos mirar en el interior del neocórtex del hablante, veríamos las conexiones y los pesos correspondientes a los modelos ocultos jerárquicos de Márkov computados por el software? Casi con seguridad, no encontraríamos una correspondencia precisa, ya que las estructuras neuronales diferirían en muchos detalles con respecto a los modelos en el interior del ordenador. Sin embargo, yo diría que debe haber una equivalencia matemática esencial con un alto grado de precisión entre la biología y nuestro intento por emularla. De otra manera, estos sistemas no funcionarían tan bien como lo hacen.

## *LISP*

LISP (LISt Processor) es un lenguaje informático definido por primera vez por el pionero de la IA John McCarthy (1927–2011) en 1958. Como su propio nombre

indica, LISP trabaja con listas. Todos los comunicados LISP son listas de elementos. Cada elemento es o bien otra lista o un "átomo", un ítem irreductible que consta o bien de un número o de un símbolo. La propia lista puede venir representada por la lista incluida en una lista, ya que LISP es capaz de funcionar recursivamente. Otra manera en la que los comunicados LISP pueden ser recursivos se da cuando una lista incluye una lista y así sucesivamente hasta que la lista original acaba siendo especificada. Como las listas pueden incluir otras listas, LISP también es capaz de realizar procesamiento jerárquico. Una lista puede estar condicionada de manera que solo se "dispare" si sus elementos son satisfechos. De esta manera, las jerarquías de dichos condicionamientos pueden ser usadas para identificar cualidades cada vez más abstractas de un patrón.

LISP hizo furor en la comunidad de la inteligencia artificial de la década de 1970 y principios de la de 1980. La idea que tenían los entusiastas de LISP de décadas pasadas era que el lenguaje reflejaba la manera en la que trabaja el cerebro humano y que todo proceso inteligente podía ser codificado de la manera más fácil y eficiente mediante LISP. Esto produjo un pequeño boom en las compañías de "inteligencia artificial" que ofertaban intérpretes LISP y productos relacionados con LISP. Sin embargo, a mediados de la década de 1980 se hizo patente que LISP no representaba un atajo para crear procesos inteligentes y la burbuja de la inversión explotó.

No obstante, resulta que los entusiastas de LISP no estaban completamente equivocados. En el fondo, todos los reconocedores de patrones del neocórtex pueden ser considerados como una expresión LISP, ya que todos conforman una lista de elementos y cada elemento puede ser otra lista. Por tanto, el neocórtex sí que está involucrado en el procesamiento de listas de naturaleza simbólica, y este procesamiento es muy similar al que tiene lugar en un programa LISP. Y no solo eso, simultáneamente procesa 300 millones de "expresiones" similares a las de LISP.

Sin embargo, en el mundo LISP había dos importantes características ausentes, una de ellas era el aprendizaje. Los programas LISP tenían que ser codificados línea a línea por programadores humanos. Se hicieron intentos por codificar programas LISP automáticamente usando diferentes métodos, pero estos no formaban parte integral del concepto del lenguaje. Por el contrario, el neocórtex se programa a sí mismo. Rellena sus "expresiones" (es decir, sus listas) con información dotada de sentido y susceptible de ser procesada proveniente de su propia experiencia y de sus propios ciclos de retroalimentación. He aquí un principio fundamental sobre el funcionamiento del neocórtex: todos sus reconocedores de patrones (es decir, todas las expresiones similares a las expresiones LISP) son capaces de rellenar sus propias listas y de conectarse a otras listas de forma tanto ascendente como descendente. La segunda diferencia es el tamaño de los parámetros. Se podría crear una variante de LISP codificada en lenguaje LISP que permitiera manejar dichos

parámetros, pero estos no formarían parte del lenguaje básico.

LISP es coherente con la filosofía original del campo de la IA, que pretendía encontrar soluciones inteligentes a problemas y codificarlas directamente en lenguajes informáticos. El primer intento para construir un método autoorganizativo que aprendiera por sí mismo a partir de la experiencia, las redes neuronales, no tuvo éxito porque no proporcionaba los medios necesarios para modificar la topología del sistema de acuerdo con el aprendizaje realizado. El modelo oculto jerárquico de Márkov proporcionó esto mediante su mecanismo de poda. Hoy, los HHMM y sus parientes matemáticos conforman la mayor parte del mundo de la IA.

Un corolario a la observación de las similitudes entre LISP y la estructura en lista del neocórtex es un argumento defendido por aquellos que defienden que el cerebro es demasiado complicado para que lo podamos comprender. Estos críticos señalan que el cerebro posee billones de conexiones y que dado que la existencia de todas y cada una de ellas viene especificada a partir del diseño original, estas constituyen el equivalente a billones de líneas de código. Tal y como hemos visto, yo calculo que en el neocórtex existen del orden de 300 millones de procesadores de patrones, o 300 millones de listas en las que cada elemento señala otra lista (o, en el nivel conceptual más bajo, un patrón básico irreductible procedente del exterior del neocórtex). Sin embargo, 300 millones sigue siendo un número razonablemente grande de expresiones LISP y de hecho es más grande que cualquier programa que haya sido escrito por humanos.

Sin embargo, tenemos que tener en cuenta que estas listas no vienen especificadas en el diseño original del sistema nervioso. El propio cerebro crea estas listas y conecta automáticamente los niveles a partir de sus propias experiencias. Este es el secreto fundamental del neocórtex. Los procesos que realizan esta autoorganización son mucho más sencillos que los 300 millones de expresiones que constituyen la capacidad del neocórtex. Dichos procesos vienen especificados en el genoma. Tal y como demostraré en el capítulo 11, la cantidad de información exclusiva contenida por el genoma, después de ser comprimida sin pérdidas, consta de 25 millones de bytes en lo que respecta al cerebro, lo cual equivale a menos de un millón de líneas de código. La complejidad algorítmica real es todavía menor que eso, ya que la mayoría de los 25 millones de bytes de la información genética pertenecen a las necesidades biológicas de las neuronas y no específicamente a su capacidad de procesar información. Sin embargo, incluso 25 millones de bytes de información sobre el diseño representan un nivel de complejidad que sí que podemos manejar.

## *Sistemas jerárquicos de memoria*

Tal y como expuse en el capítulo 3, en 2003 y 2004 Jeff Hawkins y Dileep George desarrollaron un modelo del neocórtex que incorporaba listas jerárquicas que fueron descritas en *On Intelligence*, el libro que Hawkins y Blakeslee escribieron en 2004. Una presentación más actualizada y elegante sobre este método jerárquico de memoria temporal puede encontrarse en la tesis doctoral que Dileep George presentó en 2008 (12). Numenta lo ha puesto en práctica en un sistema llamado NuPIC (Numenta Platform for Intelligent Computing) y ha desarrollado sistemas de reconocimiento de patrones y de búsqueda y procesamiento de datos para clientes tales como Forbes y Power Analytics Corporation. Después de trabajar para Numenta, George formó una nueva empresa llamada Vicarious Systems que contaba con la financiación de Founder Fund (entidad dirigida por Peter Thiel, inversor en capital riesgo que está detrás de Facebook, y por Sean Parker, el primer Presidente de Facebook) y de Good Ventures, cuyo líder es Dustin Moskovitz, cofundador de Facebook. George está haciendo públicos importantes progresos en modelización, aprendizaje y reconocimiento de información automáticos realizados mediante un número importante de jerarquías. A este sistema le llama "mecanismo cortical recursivo"* y planea desarrollar aplicaciones para imágenes médicas y robots, entre otros campos. Matemáticamente, la técnica de los modelos ocultos jerárquicos de Márkov es muy similar a estos sistemas jerárquicos de memoria, sobre todo si permitimos que el sistema HHMM organice sus propias conexiones entre los módulos de reconocimiento de patrones. Tal y como he mencionado anteriormente, los HHMMs proporcionan un elemento adicional importante, que es la modelización de la distribución esperada de la magnitud (sobre cierto continuum) de todos los input a la hora de calcular la probabilidad de la existencia del patrón que se esté considerando. Recientemente he formado una nueva empresa llamada Patterns, Inc que pretende desarrollar modelos jerárquicos neocorticales autoorganizativos que utilicen HHMMs y técnicas afines con el objetivo de comprender el lenguaje natural. Se hará mucho énfasis en la capacidad del sistema para diseñar sus propias jerarquías de forma similar a como lo hace el neocórtex biológico. El sistema que tenemos en la cabeza será capaz de leer de forma continuada una amplia gama de materiales, tales como Wikipedia y otras fuentes de conocimiento, así como escuchar cualquier cosa que se diga y observar todo lo que se escriba (si es que se le permite). El objetivo es que se convierta un amigo útil que conteste nuestras preguntas *antes* incluso de que las formulemos y que nos proporcione información útil, así como consejos, a lo largo del día.

# *La frontera móvil de la IA: ascendiendo por el escalafón de las capacidades*

1. Un discurso largo y tedioso escrito en el trivial aderezo de un pastel.

2. Un vestido llevado por un niño que quizás esté a bordo de un barco operístico.

3. Buscado por una multitud de delitos perpetrados durante doce años en los que devoró a los guerreros del rey Hroðgar; se le asignó el caso al detective Beowulf.

4. Puede significar un desarrollo mental gradual o algo que se transporta durante el embarazo.

5. El Día Nacional del Profesor y el Día del Derby de Kentucky.

6. Wordsworth dijo de ellos que planean de aquí para allá, pero que nunca vagan sin rumbo.

7. Palabra de cuatro letras para nombrar el herraje en el casco de un caballo o una caja expendedora de cartas en un casino.

8. En el tercer acto de una ópera de Verdi del año 1846 este Azote de Dios es acuchillado hasta la muerte por su amante Odabella.

> —EJEMPLO DE PREGUNTAS DEL *JEOPARDY!* QUE FUERON CONTESTADAS CORRECTAMENTE POR WATSON. LAS CONTESTACIONES SON: ARENGA DE MERENGUE, PICHI, GRENDEL, GESTAR, MAYO, LA ALONDRA, EL ZAPATO . EN LA OCTAVA PREGUNTA , WATSON RESPONDIÓ "¿QUÉ ES ATILA?" EL MODERADOR RESPONDIÓ DICIENDO: "¿PUEDES SER MÁS ESPECÍFICO?" A LO QUE WATSON RESPONDIÓ: "¿QUÉ ES ATILA EL HUNO?", QUE ES CORRECTO*.

Las técnicas del ordenador para desenmarañar las pistas del *Jeopardy!* se parecían mucho a las mías. Esa máquina asigna un cero a una palabra clave perteneciente a una pista, entonces examina su memoria (en el caso de Watson, un banco de datos de 15 terabytes de conocimiento humano) en busca de cúmulos de asociaciones que contengan esas palabras. Compara rigurosamente los resultados más probables con toda la información contextual que puede recabar: la categoría de nombre; el tipo de respuesta buscada; el momento, lugar y género que indica la pista; etc. Cuando se siente lo suficientemente "seguro" decide hacer sonar el timbre. Todo esto no es más que un instante del proceso intuitivo de un jugador humano del *Jeopardy!* Sin embargo, estaba convencido de que debajo de la capucha mi cerebro estaba haciendo más o menos la misma cosa.

> —KEN JENNINGS, CAMPEÓN HUMANO DE *JEOPARDY!* QUE PERDIÓ CONTRA WATSON.

Yo por mi parte doy la bienvenida a nuestros nuevos jefes supremos robóticos.

> —KEN JENNINGS PARAFRASEANDO A *LOS SIMPSONS* DESPUÉS DE PERDER CONTRA WATSON

Oh, dios mío. [Watson] es más inteligente que el jugador medio de *Jeopardy!* a la hora de responder las preguntas de *Jeopardy!* Es algo asombrosamente inteligente.

> —SEBASTIAN THRUN, EXDIRECTOR DEL LABORATORIO DE IA DE STANFORD

Watson no entiende nada. Es solo una apisonadora más grande.

> —NOAM CHOMSKY

Estamos rodeados de inteligencia artificial y ya no tenemos la mano puesta sobre el enchufe. El simple hecho de contactar con alguien a través de un mensaje de texto, de un email o de una llamada de teléfono móvil implica el uso de algoritmos inteligentes que enrutan la información. Casi todos los productos a nuestro alcance están originariamente diseñados mediante la colaboración entre la inteligencia humana y la artificial, y después son construidos en fábricas automatizadas. Si mañana todos los sistemas de IA decidieran hacer huelga, nuestra civilización se vería incapacitada. No podríamos sacar dinero del banco. De hecho, nuestro dinero desaparecería. Las comunicaciones, los transportes y las manufacturas se detendrían. Por fortuna, nuestras inteligentes máquinas no son todavía lo suficientemente inteligentes como para organizar una conspiración así.

Lo nuevo de la IA de hoy es lo visceralmente impresionante que es la naturaleza de los ejemplos asequibles por el público. Por ejemplo, tomemos en consideración los coches autoconducidos de Google, que mientras este libro está siendo escrito han recorrido más de 200 000 millas en ciudades y pueblos, una tecnología que conllevará un descenso significativo de los accidentes, aumentará la capacidad de las carreteras, liberará a los humanos de la tarea de conducir y que conllevará muchos otros beneficios. Los coches sin conductor ya son legales en las carreteras públicas de Nevada, aunque bajo algunas restricciones. Por otra parte, el uso generalizado de este tipo de coches en el mundo no se espera que se produzca hasta finales de esta década. En los coches ya se está instalando tecnología que de forma inteligente vigila la carretera y avisa al conductor de peligros inminentes. En parte, una tecnología así se basa en el éxito a la hora de modelizar el procesamiento visual del cerebro llevado a cabo por Tomaso Poggio en el MIT. Bajo el nombre de MobilEye, fue desarrollado por Amnon Shashua, un antiguo estudiante de postdoctorado de Poggio. Es capaz de alertar al conductor de peligros tales como una colisión inminente o un niño que cruza corriendo delante del coche. Recientemente ha sido instalado por fabricantes de coches tales como Volvo y BMW.

Por varias razones, en esta sección del libro me centraré en tecnologías del lenguaje. No es sorprendente que la naturaleza jerárquica del lenguaje refleje fielmente la naturaleza jerárquica de nuestro pensamiento. El lenguaje hablado fue nuestra primera tecnología, y el lenguaje escrito fue la segunda. Mi propio trabajo en el campo de la inteligencia artificial, tal y como ha quedado demostrado en este capítulo, se ha centrado especialmente en el lenguaje. En último término, el dominio del lenguaje ofrece posibilidades muy poderosas. Watson ya ha leído cientos de millones de páginas en la web, y ya domina el conocimiento contenido en dichos documentos. En último término, las máquinas serán capaces de dominar todo el conocimiento contenido en la web, que es básicamente todo el conocimiento de

nuestra civilización hombre-máquina.

El matemático inglés Alan Turing (1912–1954) basó su epónimo test en la capacidad de un ordenador para conversar en lenguaje natural usando mensajes de texto (13). Turing creía que toda la inteligencia humana se plasmaba y venía representada por el lenguaje, y que ninguna máquina podría pasar un test de Turing mediante simples trucos del lenguaje. Aunque el test de Turing es un juego que involucra lenguaje escrito, Turing creía que la única manera en la que un ordenador podría pasarlo sería poseyendo un nivel de inteligencia equivalente al humano. Por su parte, los críticos de esta idea han señalado que un verdadero test de nivel de inteligencia humano también debería incluir el dominio de información visual y auditiva (14). Dado que muchos de mis propios proyectos de IA conllevan enseñar a los ordenadores para que dominen una información sensorial como el habla humana, las formas de las letras y los sonidos musicales, podría esperarse que yo también abogara por la inclusión de estas formas de información en un verdadero test de inteligencia. Sin embargo, estoy de acuerdo con la idea original de Turing de que la versión del test de Turing basada solamente en el texto es suficiente. El añadir al test input u output, ya sea visual o auditivo, no haría que pasarlo fuese más difícil.

No es necesario ser un experto en IA para emocionarse con el comportamiento de Watson en *Jeopardy!* Aunque poseo una comprensión razonable sobre la metodología usada en varios de sus subsistemas fundamentales, eso no disminuye mi reacción emocional al verle *(¿a él?)* en acción. Incluso un conocimiento perfecto de cómo funcionan todos los sistemas que lo componen, conocimiento que por cierto nadie posee, no ayudaría a predecir la manera en que Watson reacciona ante una determinada situación. Contiene cientos de subsistemas que interactúan entre sí, y cada uno de ellos sopesa millones de posibles hipótesis al mismo tiempo, de manera que predecir el resultado es imposible. Un análisis exhaustivo a posteriori de las deliberaciones que Watson realiza solamente durante una indagación de tres segundos le llevaría a un humano varios siglos.

Continuando con mi propia historia, a finales de la década de 1980 y en la década de 1990 empezamos a trabajar en la comprensión del lenguaje natural en dominios limitados. Se podía hablar a uno de nuestros productos llamado Kurzweil Voice sobre cualquier cosa que se quisiera, siempre y cuando tuviera que ver con la edición de documentos. (Por ejemplo, “mover hasta aquí el tercer párrafo de la página anterior”). Funcionaba bastante bien en este limitado pero útil dominio. También creamos sistemas con conocimientos médicos, de manera que los médicos pudieran dictar los informes del paciente. Poseía suficiente conocimiento de campos tales como la radiología y la patología como para poder preguntar al médico en el caso de que algo en el informe pareciera poco claro, y como para poder guiar al facultativo a lo largo del proceso de realización del informe. Estos sistemas médicos han evolucionado mediante Nuance hasta convertirse en negocios de miles de

millones de dólares.

La comprensión del lenguaje natural, sobre todo como extensión del reconocimiento automático del habla, es ya un fenómeno dominante. Mientras se escribe este libro, Siri, el asistente personal automatizado del iPhone 4S, ha revolucionado el mundo de la computación portátil. A Siri se le puede pedir que haga cualquier cosa que un *smartphone* con amor propio debería ser capaz de hacer (por ejemplo, "¿dónde puedo conseguir comida india por aquí cerca?" o "envía un mensaje de texto a mi mujer diciendo que estoy de camino" o "¿qué opina la gente de la nueva película de Brad Pitt?"), y la mayor parte de las veces Siri cumplirá su tarea. Siri puede mantener una pequeña conversación banal. Si le preguntas cuál es el sentido de la vida, responderá que "42", que los fans de *La guía del autoestopista galáctico** reconocerán como "la respuesta a la pregunta última sobre la vida, el universo y el todo". Las preguntas de razonamiento (incluyendo la que versa sobre el sentido de la vida) son contestadas por Wolfram Alpha, está descrito en la página 163. Existe todo un mundo de "chatbots" que no hacen otra cosa más que participar en charlas irrelevantes. Si usted desea hablar con nuestro chatbot llamado Ramona, visite nuestra web de KurzweilAI.net y cliquee en "Chat with Ramona".

Algunas personas han venido a mí para criticar la incapacidad de Siri a la hora de atender ciertas peticiones. Sin embargo, me suelo dar cuenta de que son las mismas personas que una y otra vez también se quejan de los que ofertan servicios humanos. A veces les propongo que lo intentemos juntos y a menudo funciona mejor de lo esperado. Las quejas me recuerdan a la historia del perro que juega al ajedrez. Ante un interlocutor incrédulo el dueño del perro contesta: "sí, es verdad, juega al ajedrez, pero flaquea en los finales". Además, a Siri ya le están saliendo competidores muy fuertes tales como Google Voice Search.

El hecho de que el público en general ya esté manteniendo conversaciones en lenguaje natural hablado con sus ordenadores portátiles marca el comienzo de una nueva era. Debido a sus limitaciones, es típico que la gente desestime la importancia de las tecnologías de primera generación. Unos años después, cuando la tecnología funciona bien, la gente sigue desestimándola, ya que… bueno… ya no es nueva. Dicho esto, para ser un producto de primera generación Siri funciona sorprendentemente bien, y está claro que esta categoría de productos no va a hacer más que mejorar.

Siri utiliza tecnologías de reconocimiento del habla basadas en HHMM procedentes de Nuance. Las extensiones del lenguaje natural fueron desarrolladas por primera vez por el proyecto financiado por DARPA llamado "CALO" (15). Siri ha ido mejorando mediante las tecnologías de lenguaje natural de Nuance, y Nuance ofrece ya una tecnología muy similar llamada Dragon Go! (16)

Los métodos usados para entender el lenguaje natural son muy similares a los

modelos ocultos jerárquicos de Márkov, y de hecho el propio HHMM es de uso común. Aunque algunos de estos sistemas no son catalogados específicamente como usuarios de HHMM o directamente de HHMM, las matemáticas utilizadas son virtualmente idénticas. Todos involucran jerarquías de secuencias lineales en las que cada elemento tiene un peso, las conexiones se autoadaptan y el sistema general se autoorganiza según los datos del aprendizaje. Por lo general, el aprendizaje se extiende durante el propio uso del sistema. Esta estrategia se corresponde con la estructura jerárquica del lenguaje natural, se trata tan solo de una extensión natural hacia arriba por la escalera conceptual que va desde las partes del habla hasta las palabras, de ahí hasta las frases, y desde las frases hasta las estructuras semánticas. Por eso, tendría sentido ejecutar un algoritmo genético sobre los parámetros que controlan justamente el algoritmo del aprendizaje de esta clase de sistemas de aprendizaje jerárquicos y determinar los detalles algorítmicos óptimos.

Durante la última década se ha producido un cambio en la forma en la que estas estructuras jerárquicas son creadas. En 1984, Douglas Lenat (nacido en 1950) inauguró el ambicioso proyecto Cyc (cuyo nombre proviene de enCYClopedic), que buscaba crear reglas que codificaran los conocimientos de "sentido común" del día a día. Las reglas fueron organizadas según una inmensa jerarquía y cada regla conllevaba una nueva secuencia lineal de estados. Por ejemplo, una regla Cyc podría sostener que un perro tiene cara. Entonces, Cyc puede enlazar con las reglas generales sobre la estructura de las caras: una cara tiene dos ojos, una nariz, una boca, etc. No se necesita tener un conjunto de reglas correspondientes a la cara del perro y luego otro correspondiente a la cara del gato, aunque por supuesto es posible que se quisieran añadir reglas adicionales para los aspectos en los que las caras de los perros se diferencian de las caras de los gatos. El sistema también incluye un mecanismo de inferencia. Así, si tenemos reglas que sostienen que un cocker spaniel es un perro, que los perros son animales y que los animales comen comida, y asimismo preguntáramos al mecanismo de inferencia si un cocker spaniel come, el sistema respondería que sí, los cocker spaniels comen comida. Durante los siguientes 25 años, mediante miles de años de trabajo humano, se escribieron y probaron más de un millón de dichas reglas. Curiosamente, el lenguaje para escribir reglas Cyc, llamado CycL, es casi idéntico a LISP.

Mientras tanto, una escuela de pensamiento opuesta sostuvo que la mejor estrategia para comprender el lenguaje natural y para crear sistemas inteligentes en general era a través del aprendizaje automático surgido de la exposición a un gran número de casos relacionados con los fenómenos que el sistema estuviera intentando dominar. Un ejemplo poderoso de un sistema así es Google Translate, que puede traducir a y desde 50 idiomas. Eso significa 2 500 direcciones de traducción diferentes, aunque para la mayoría de combinaciones en vez de traducir el lenguaje 1 directamente al lenguaje 2 se traduce el lenguaje 1 al inglés y luego del inglés al

lenguaje 2. Esto reduce el número de traductores que Google necesitó construir a solo 98 (más un número limitado de emparejamientos no ingleses para los que existe una traducción directa). Los traductores de Google no usan reglas gramaticales. En su lugar, crean una enorme base de datos para cada emparejamiento de traducciones habituales que se basa en amplios cuerpos de documentos traducidos de idioma a idioma y que hacen las veces de "piedra roseta". Para las 6 lenguas que conforman las lenguas oficiales de la Naciones Unidas, Google ha utilizado documentos de las Naciones Unidas tal y como fueron publicados en las 6 lenguas. Para idiomas menos comunes se han utilizado otras fuentes.

A menudo los resultados son sorprendentes. DARPA organiza competiciones anuales para determinar los mejores sistemas de traducción automatizados según diferentes emparejamientos de idiomas, y Google Translate suele ganar en ciertos emparejamientos derrotando a sistemas creados directamente por lingüistas humanos.

Dos perspectivas fundamentales han influenciado profundamente el campo de la comprensión del lenguaje natural durante las últimas dos décadas. La primera tiene que ver con las jerarquías. Aunque la estrategia de Google comenzó con la asociación de secuencias planas de palabras desde una lengua a otra, la inherente naturaleza jerárquica del lenguaje se ha deslizado en su funcionamiento. Los sistemas que de forma metódica incorporan aprendizaje jerárquico (tal y como pasa con los modelos ocultos jerárquicos de Márkov) han demostrado un rendimiento significativamente mejor. Sin embargo, la construcción de dichos sistemas no es un proceso excesivamente automático. Al igual que los humanos se ven obligados a aprender aproximadamente una jerarquía conceptual cada vez, lo mismo les pasa a los sistemas informáticos, de manera que el proceso de aprendizaje tiene que organizarse cuidadosamente.

La otra perspectiva indica que las reglas hechas a mano funcionan bien para un núcleo basado en los conocimientos generales básicos. Para la traducción de pasajes cortos, esta estrategia suele reportar resultados más precisos. Por ejemplo, DARPA ha calificado a los traductores de chino a inglés basados en reglas con una mejor nota que la de Google Translate para el caso de pasajes cortos. En lo que respecta a los denominados flecos de una lengua*, que hacen referencia a los millones de expresiones infrecuentes y a los conceptos que se utilizan en ellas, la precisión de los sistemas basados en reglas se acerca a una asíntota inaceptablemente baja. Si se representan en un mismo gráfico comparativo la precisión en la comprensión del lenguaje natural y la cantidad de datos de entrenamiento analizados, los sistemas basados en reglas muestran inicialmente un rendimiento mayor, pero que se compensa con una precisión bastante baja de más o menos el 70%. Muy por el contrario, los sistemas estadísticos pueden alcanzar una precisión superior al 90%,

pero para ello necesitan una gran cantidad de datos.

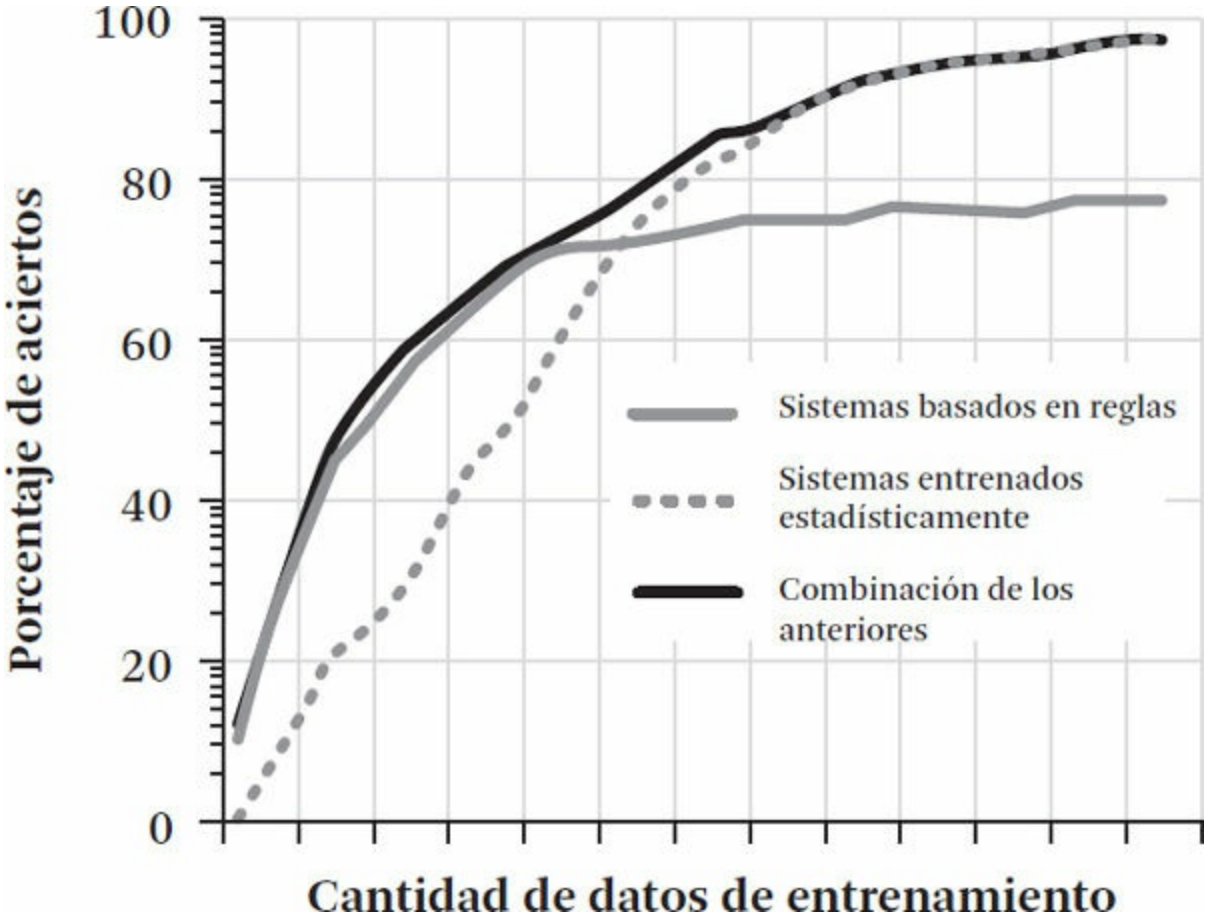
A menudo se necesita una combinación de por lo menos un rendimiento moderado con respecto a una pequeña cantidad de datos de entrenamiento para luego poder alcanzar precisiones altas con una cantidad de datos que sea notablemente mayor. Alcanzar rendimientos moderados permite introducir rápidamente un sistema sobre el terreno para luego reunir datos de entrenamiento automáticamente a medida que la gente lo utiliza. De esta manera, una gran cantidad de aprendizaje puede tener lugar al mismo tiempo que el sistema está siendo utilizado, y además su precisión aumentará. El aprendizaje estadístico tiene que ser completamente jerárquico para que refleje la naturaleza del lenguaje, cosa que también refleja el modo en el que funciona el cerebro humano.

Así es como funcionan Siri y Dragon Go!, usando reglas para los fenómenos más comunes y predecibles, y después aprendiendo los "flecos" del lenguaje dejándolos en manos de usuarios reales. Cuando el equipo de Cyc se dio cuenta de que habían alcanzado un techo en el rendimiento de las reglas codificadas a mano, también ellos adoptaron esta estrategia. Las reglas codificadas a mano proporcionan dos funciones esenciales. Ofrecen una precisión inicial adecuada, de manera que un sistema de prueba que mejorará automáticamente pueda ser ampliamente utilizado. En segundo lugar, proporcionan una base sólida para los niveles más bajos de la jerarquía conceptual, de manera que el aprendizaje automatizado pueda comenzar a aprender niveles conceptuales más elevados.

Como ya he dicho anteriormente, Watson representa un ejemplo particularmente impresionante de la estrategia que combina reglas codificadas a mano con el aprendizaje jerárquico estático. IBM combinó una serie de programas de lenguaje natural punteros para crear un sistema que podía jugar al juego de lenguaje natural llamado *Jeopardy!* Entre el 14 y el 16 de febrero de 2011, Watson compitió contra los dos mejores jugadores humanos: Brad Rutter, que había ganado más dinero que nadie en el concurso de preguntas y respuestas, y Ken Jennings, que anteriormente había ganado *Jeopardy!* durante un periodo record de 75 días.

Como nota contextual diré que en mi primer libro, *The Age of Intelligent Machines*, escrito a mediados de la década de 1980, predije que un ordenador ganaría el campeonato mundial de ajedrez hacia 1998. También predije que cuando eso pasara, una de tres: o bien rebajaríamos nuestra opinión sobre la inteligencia humana, o bien mejoraríamos nuestra opinión sobre la inteligencia de las máquinas, o bien disminuiríamos la importancia que le damos al ajedrez y, tomando a la historia como referencia, el juego del ajedrez disminuiría en popularidad. Las dos primeras cosas pasaron el 1997. El superordenador de IBM llamado Deep Blue derrotó al vigente campeón mundial, Garry Kasparov. Inmediatamente después empezamos a escuchar argumentos que decían que era de esperar que un ordenador ganara al ajedrez, ya que los ordenadores son máquinas lógicas y el ajedrez, después

de todo, es un juego de lógica. Así, la victoria de Deep Blue no fue juzgada ni de sorprendente, ni de significativa, y muchos de los críticos defendieron que los ordenadores nunca dominarían las sutilezas del lenguaje humano, incluyendo metáforas, símiles, juegos de palabras, doblesentidos y chistes.




**La precisión de los sistemas de comprensión del lenguaje natural en función de la cantidad de datos de entrenamiento. La mejor estrategia es la de combinar reglas para el "núcleo" del lenguaje y una estrategia basada en datos para los "flecos" del lenguaje.**

Por lo menos esta es una razón por la que Watson representa un hito tan importante: *Jeopardy!* significa precisamente dicha tarea de sofisticación y reto lingüístico. Las típicas preguntas de *Jeopardy!* incluyen muchas de estas extravagancias del lenguaje humano. De lo que quizá muchos analistas no se percaten es que Watson no solo tuvo que dominar el lenguaje de las desconocidas e intrincadas preguntas, sino que además la mayor parte de su conocimiento no había sido codificado a mano. De hecho, dicho conocimiento lo obtuvo de leer 200 millones de páginas procedentes de documentos en lenguaje natural, incluyendo enciclopedias enteras entre las que se encontraba Wikipedia y abarcando 4 billones de bytes de conocimiento basado en el lenguaje. Tal y como los lectores de este libro saben muy bien, Wikipedia no está escrita ni en LISP ni en CycL, sino en frases naturales que poseen todas las ambigüedades y complicaciones inherentes al lenguaje. Para responder una pregunta, Watson tenía que tener en cuenta 4 billones

de caracteres contenidos en su material de referencia. (Soy consciente de que las preguntas de *Jeopardy!* son respuestas para las que hay que formular una pregunta, pero esto no es más que un tecnicismo, en último término se trata de preguntas). Si Watson puede comprender y responder cuestiones contenidas en 200 millones de páginas (¡en solo tres segundos!), no hay nada que impida que sistemas similares lean el resto de miles de millones de documentos que se encuentran en la web. De hecho, dicho intento ya se está realizando.

Cuando, entre la década de 1970 hasta la de 1990, desarrollábamos sistemas de reconocimiento de caracteres y del habla, así como sistemas primitivos para la comprensión del lenguaje natural, utilizábamos como metodología la introducción de un "gestor experto"*. Desarrollábamos múltiples sistemas que hacían lo mismo, pero que incorporaban estrategias ligeramente diferentes. Algunas de estas diferencias eran nimias, como por ejemplo variaciones en los parámetros de control de los procesos matemáticos pertenecientes al algoritmo de aprendizaje. Otras variaciones eran fundamentales, como por ejemplo la introducción de sistemas basados en reglas en lugar de sistemas jerárquicos de aprendizaje estadístico. El propio gestor experto era un programa de software programado para aprender los puntos fuertes y débiles de estos sistemas mediante el examen de sus rendimientos en situaciones de la vida real. Se basaba en la idea de que dichos puntos fuertes eran ortogonales**, es decir, que un sistema tendía a ser mejor en el punto en el que otro tendía a hacerlo peor. De hecho, el rendimiento general de los sistemas que se combinaban con el gestor experto previamente entrenado, era mucho mejor que cualquiera de los sistemas individuales.

Watson funciona de la misma manera. Mediante una estructura llamada UIMA (*Unstructured Information Management Architecture*) , Watson hace uso de literalmente cientos de sistemas diferentes (muchos de los componentes individuales del lenguaje de Watson son los mismos que se utilizan en los sistemas de comprensión del lenguaje natural de venta al público), la totalidad de los cuales intenta, o bien encontrar directamente la respuesta a la pregunta de *Jeopardy!* o por lo menos proporcionar algún tipo de aclaración de la pregunta en cuestión. Básicamente, UIMA actúa como el gestor experto que combina inteligentemente los resultados de los sistemas independientes. UIMA va bastante más allá que sistemas anteriores, tales como el sistema que desarrollamos en la compañía predecesora de Nuance, ya que sus sistemas individuales pueden contribuir a la obtención de resultados sin tener que encontrar necesariamente la respuesta definitiva. Basta con que un subsistema ayude a reducir el marco de posibles soluciones. UIMA también puede calcular su grado de confianza en la respuesta final. También el cerebro humano hace esto. Probablemente estemos muy seguros de nuestra respuesta si nos preguntan por el nombre de pila de nuestra madre, pero probablemente lo estemos menos si nos preguntan por el nombre de alguien con quien nos encontramos hace un

año por casualidad.

Así, en vez de optar por una única y elegante estrategia para comprender el problema del lenguaje inherente a *Jeopardy!*, los científicos de IBM combinaron todos los módulos de comprensión del lenguaje disponibles a los que tuvieron acceso. Algunos utilizan modelos ocultos jerárquicos de Márkov, algunos usan variantes matemáticas de HHMM, otros utilizan estrategias basadas en reglas para codificar directamente un núcleo de reglas en las que se puede confiar. Además, UIMA evalúa el rendimiento de cada sistema durante su uso real y lo combina de forma óptima. En el debate público existe cierto grado de confusión sobre Watson, ya que los científicos de IBM que lo crearon suelen centrarse en UIMA, el gestor experto que ellos mismos crearon. Esto hace que algunos analistas digan que Watson no posee una comprensión real del lenguaje, ya que es difícil identificar dónde se encuentra dicha comprensión. Aunque los parámetros de funcionamiento de UIMA también incorporan el aprendizaje a partir de la experiencia, la "comprensión" que Watson posee del lenguaje no puede encontrase exclusivamente en UIMA, sino que está distribuido por todos sus componentes, incluyendo los módulos autoorganizativos del lenguaje que usan métodos similares a HHMM.

Otra parte distinta de la tecnología de Watson utiliza la estimación de la confianza en las respuestas realizada por UIMA para determinar el modo de realizar las apuestas del *Jeopardy!* Aunque el sistema Watson está específicamente optimizado para jugar a este juego en concreto, el núcleo de su tecnología lingüística y su tecnología de búsqueda de conocimiento pueden ser fácilmente adaptados para realizar otras tareas. Se podría pensar que conocimientos profesionales comúnmente menos compartidos, tales como el conocimiento médico, podrían ser más difíciles de dominar que el conocimiento "de cultura general" necesario para jugar al *Jeopardy!* De hecho, lo cierto es lo contrario. El conocimiento profesional tiende a encontrarse más organizado y estructurado, sin el alto grado de ambigüedad de la cultura general. Así, el conocimiento profesional es muy apropiado para estas técnicas de comprensión precisa del lenguaje natural. Tal y como ya he mencionado, actualmente IBM está trabajando junto con Nuance para adaptar la tecnología Watson al campo de la medicina.

La conversación que se produce cuando Watson juega al *Jeopardy!* es muy breve: se formula una cuestión y Watson da una respuesta. (Repito, técnicamente, Watson formula una pregunta para responder a una contestación). No se produce una conversación en la que sea necesario estar al tanto de todas las declaraciones hechas por cada uno de los participantes. (De hecho, hasta cierto punto esto es lo que hace Siri, ya que si le pide que envíe un mensaje a su mujer, le pedirá que la identifique pero en ocasiones posteriores se acordará de quién es ella). Estar al tanto de toda la información contenida en una conversación, una tarea claramente necesaria para

pasar el test de Turing, es un importante requisito adicional, pero no se trata de algo que fundamentalmente sea más complicado de lo que ya realiza Watson. Después de todo, Watson ha leído cientos de millones de páginas informativas que obviamente incluyen muchas historias, de manera que sí que es capaz de seguir el hilo a través de complicadas secuencias de eventos. Por tanto, debería poder seguir sus propias conversaciones y tomarlas en consideración en respuestas subsiguientes.

Otra limitación del juego del *Jeopardy!* es que por lo general las respuestas son breves. Por ejemplo, no plantea el tipo de preguntas que interrogan a los concursantes sobre el nombre de los cinco temas principales de "Historia de dos ciudades"*. Teniendo en cuenta que puede encontrar documentos que debaten sobre los temas de esta novela, una versión de Watson modificada a tal efecto debería ser capaz de responder a esta pregunta. Proponer él mismo dichos temas a partir de la mera lectura del libro y no limitarse a copiar la reflexión de otros pensadores (aunque quizá con sus propias palabras) es otra cuestión diferente. Realizar algo así constituye una tarea de un nivel más alto del que Watson, a día de hoy, es capaz de alcanzar. Se trataría de lo que yo llamo una tarea a nivel del Test de Turing. (Dicho esto, he de decir que la mayoría de los humanos tampoco plantean por sí mismos pensamientos originales, si no que copian las ideas de sus coetáneos y líderes de opinión). En cualquier caso, estamos en 2012, no en 2029, de manera que todavía no hay que aspirar a un nivel de inteligencia correspondiente al test de Turing. No obstante, he de señalar que la evaluación de respuestas a preguntas tales como la de encontrar las ideas fundamentales de una novela no es una tarea sencilla. Si a alguien se le preguntara quién firmó la Declaración de Independencia de los EE.UU., se podría determinar si la respuesta dada es verdadera o falsa. La validez de respuestas a cuestiones de nivel más alto, como por ejemplo la descripción de los temas contenidos en una creación literaria es algo mucho más difícilmente definible.

Es de reseñar que aunque sus capacidades lingüísticas están un tanto por debajo de las de un humano culto, Watson fue capaz de derrotar a los dos mejores jugadores de *Jeopardy!* del mundo. Lo logró gracias a su capacidad para combinar sus dotes lingüísticas y comprensión de conocimientos con una perfecta recopilación de recuerdos altamente precisos propios de una máquina. Esta es la razón por la cual ya hemos confiado a las máquinas nuestros recuerdos personales, sociales e históricos.

Aunque no estoy en disposición de adelantar la fecha de mi predicción sobre el paso del test de Turing por parte de una máquina en el año 2029, el progreso conseguido en sistemas como Watson debería darnos confianza en que el nacimiento de una IA de nivel Turing está a la vuelta de la esquina. Si se creara una versión de Watson optimizada para el test de Turing, seguramente se estaría muy cerca de lograrlo.

El filósofo norteamericano John Searle (nacido en 1932) arguyó recientemente que Watson no es capaz de pensar. Haciendo referencia al experimento mental de su

“habitación china” (tema que discutiré más en profundidad en el capítulo 11), Searle defiende que Watson se limita a manipular símbolos y que no comprende el significado de dichos símbolos. En realidad, Searle no describe a Watson de forma precisa, ya que la comprensión del lenguaje que posee este está basada en procesos estadísticos jerárquicos, no en la manipulación de símbolos. La caracterización de Searle solo podría ser precisa si cada paso en los procesos autoorganizativos de Watson pudiera ser considerado una “manipulación de símbolos”. Pero si ese fuera el caso, entonces el cerebro humano tampoco podría ser considerado como capaz de llevar a cabo el pensamiento.

Es gracioso a la vez que irónico ver cómo los analistas critican a Watson por *solamente* llevar a cabo un análisis estadístico del lenguaje en contraposición a la posesión de la “verdadera” comprensión del lenguaje que poseen los humanos. El análisis estadístico y jerárquico es exactamente lo que lleva a cabo el cerebro al resolver hipótesis múltiples basadas en la inferencia estadística (cosa que de hecho se produce en todos los niveles de la jerarquía neocortical). Tanto Watson como el cerebro humano aprenden y responden según una estrategia similar a la comprensión jerárquica. En muchos aspectos, el conocimiento de Watson es mucho más extenso que el humano. Ningún humano puede decir que se ha aprendido Wikipedia al completo, cosa que es solo una parte de la base de conocimientos de Watson. Por el contrario, a día de hoy un humano puede dominar más niveles conceptuales que Watson, pero esto no supone una diferencia que vaya a ser duradera.

Un importante sistema que demuestra el poderío de la informática aplicada al conocimiento organizado es Wolfram Alpha, un mecanismo de respuestas (en contraposición a los mecanismos de búsqueda o buscadores) desarrollado por el matemático y científico británico Dr. Wolfram (nacido en 1959) y sus colegas del *Wolfram Research*.

Por ejemplo, si a Wolfram Alpha (en WolframAlpha.com) se le pregunta “¿cuántos números primos hay en un millón?”, responderá que “78 498”. No es que busque la respuesta, es que la calcula, y si se sigue el proceso de dicha respuesta también proporciona las ecuaciones que ha utilizado. Si se intentara conseguir una contestación mediante un buscador convencional, este nos dirigiría a los enlaces en los que pudiéramos encontrar los algoritmos pertinentes. Entonces habría que introducir dichas fórmulas en un sistema como “Mathematica” (también desarrollado por el Dr. Wolfram). Sin embargo, es obvio que esto conllevaría mucho más trabajo (y comprensión) que una simple pregunta a Alpha.

De hecho, Alpha consta de 15 millones de líneas de código pertenecientes a Mathematica. Lo que hace literalmente Alpha es calcular la respuesta a partir de aproximadamente 10 billones de bytes de datos que han sido cuidadosamente seleccionados por el personal de *Wolfram Research*. Se le pueden hacer un amplio

rango de preguntas fácticas, como por ejemplo "¿cuál es el país con el mayor PIB per cápita?" (Respuesta: Mónaco con $ 212 000 por persona), o "¿cuántos años tiene Stephen Wolfram?" (Respuesta: 52 años, 9 meses y 2 días en el día en que me encuentro escribiendo esto). Tal y como he mencionado, Alpha es utilizado como parte integrante del programa de Apple llamado Siri. Si a Siri se le plantea una cuestión fáctica, esta es cedida a Alpha para que se encargue de ella. Alpha también se ocupa de algunas de las búsquedas que se hacen mediante el buscador de Microsoft llamado Bing.

En una reciente entrada de blog, el Dr. Wolfram anunció que actualmente Alpha proporciona las respuestas correctas en el 90% de las veces (17). También anunció un descenso exponencial en el margen de error, que cada 18 meses se reduce a la mitad. Se trata de un sistema impresionante que utiliza métodos artesanos y datos comprobados manualmente. Para empezar, esto da muestra de por qué creamos los ordenadores. A medida que descubrimos y recopilamos métodos científicos y matemáticos, los ordenadores demuestran ser mucho mejores que la inteligencia humana a la hora de implementarlos. La mayoría de los métodos científicos conocidos han sido codificados en Alpha, e incluyen datos permanentemente actualizados sobre temas que van desde la economía hasta la física. Durante un conversación privada que mantuve con el Dr. Wolfram, calculó que los métodos autoorganizativos como los usados por Watson suelen alcanzar una precisión de alrededor del 80% cuando funcionan bien. Alpha, señaló, está alcanzando una precisión de alrededor del 90%. Por supuesto, en ambos índices de precisión existe un cierto grado de autoselección, ya que los usuarios (como por ejemplo yo mismo) han aprendido qué tipo de preguntas se le dan bien a Alpha, y un factor similar se aplica a los métodos autoorganizativos. 80% parece ser una estimación razonable sobre el nivel de precisión de Watson en las preguntas de *Jeopardy!*, pero fue suficiente para derrotar a los mejores humanos.

En mi opinión, los métodos autoorganizativos como los que he descrito en la teoría de la mente basada en el reconocimiento de patrones son necesarios para comprender las complicadas y a menudo ambiguas jerarquías que nos encontramos en los fenómenos de la vida real, el lenguaje humano incluido. Una combinación ideal para un sistema sólidamente inteligente sería una combinación de inteligencia jerárquica basada en la PRTM (según la cual yo defiendo que funciona el cerebro humano) y una precisa codificación de datos y conocimiento científico. Esto describe la esencia de lo que es un humano con un ordenador. Durante los próximos años, vamos a mejorar ambos polos de la inteligencia. En lo que respecta a nuestra inteligencia biológica, y pese a que nuestro neocórtex posee un nivel de plasticidad importante, su arquitectura básica se ve limitada por sus barreras físicas. La incorporación de neocórtex adicional a nuestras frentes supuso una importante innovación evolutiva, pero ahora no nos es posible multiplicar fácilmente el tamaño

de nuestros lóbulos frontales por mil, o siquiera aumentarlo un 10%. Es decir, no podemos lograrlo biológicamente, pero tecnológicamente es exactamente lo que haremos.

## *Una estrategia para crear una mente*

> En nuestro cerebro hay miles de millones de neuronas, pero ¿qué son las neuronas? Tan solo células. Hasta que las conexiones entre neuronas no se producen, nuestro cerebro no posee ningún conocimiento. Por lo que sabemos, todo lo que somos proviene de la manera en la que nuestras neuronas están conectadas.
>
> —TIM BERNERS-LEE

Usemos las observaciones que he expuesto más arriba para empezar a construir un cerebro. Empezaremos por construir un reconocedor de patrones que posea los atributos necesarios. Después haremos tantas copias del reconocedor como nos lo permitan nuestras capacidades de memoria y de cálculo. Cada reconocedor calcula la probabilidad de que su patrón haya sido reconocido. Al hacerlo, toma en consideración la magnitud observada de cada input (según el continuum apropiado) y la pone en relación con los parámetros de tamaño aprendidos y con los parámetros de variabilidad del tamaño asociados con cada input. El reconocedor hace que su axón simulado se dispare en el caso de que la probabilidad calculada supere un determinado umbral. Este umbral y los parámetros que controlan el cálculo de la probabilidad del patrón se encuentran entre los parámetros que optimizaremos mediante un algoritmo genético. Como no es necesario que todos los inputs estén activos para que un patrón sea reconocido, esto proporciona un reconocimiento autoasociativo (es decir, el reconocimiento de un patrón basándose en un patrón que solamente se encuentra parcialmente presente). También permitimos que haya señales inhibidoras (señales que indican que el patrón es menos probable).

El reconocimiento del patrón envía una señal activa por el axón simulado de su reconocedor de patrones. A su vez, este axón es conectado a uno o más reconocedores de patrones en el siguiente nivel conceptual superior. Todos los reconocedores de patrones conectados al siguiente nivel conceptual superior aceptan a este patrón como uno de sus inputs. Cada reconocedor de patrones también envía señales hasta los reconocedores de patrones en los niveles conceptuales inferiores en los que se ha reconocido la mayor parte de un patrón. Esto indica que el resto del patrón es “esperado”. Cada reconocedor de patrones posee uno o más de estos canales de input cuya señal es esperada. Cuando de esta manera se recibe una señal esperada, el umbral de reconocimiento de este reconocedor de patrones es rebajado (se hace más sencillo).

Los reconocedores de patrones son los responsables de “cablearse” a sí mismos

para unirse a otros reconocedores de patrones por encima y por debajo de la jerarquía conceptual. Téngase en cuenta que todo "cableado" de una implementación de software funciona por medio de enlaces virtuales (los cuales, al igual que los enlaces web, son básicamente apuntadores de memoria) y no por medio de cables reales. Así, este sistema es mucho más flexible que el del cerebro biológico. En un cerebro humano, los nuevos patrones tienen que ser asignados a un reconocedor de patrones físico, y las conexiones nuevas tienen que producirse mediante un enlace real axón-dendrita. Normalmente esto significa tomar una conexión física ya existente que sea aproximadamente igual a la que se necesita y luego desarrollar las necesarias extensiones del axón y de la dendrita para realizar la conexión completa.

Otra técnica utilizada por los cerebros biológicos de los mamíferos es la de comenzar con un gran número de posibles conexiones y luego podar las conexiones neuronales que no se usen. Si, para aprender materiales más recientes, un neocórtex biológico reasigna reconocedores de patrones corticales que ya han aprendido patrones más antiguos, entonces las conexiones tienen que ser físicamente reconfiguradas. Repito, estas tareas son mucho más sencillas en las implementaciones de software. En ellas, nos limitamos a asignar nuevas localizaciones de memoria a un nuevo reconocedor de patrones y usamos los enlaces de memoria para las conexiones. Si el neocórtex digital deseara reasignar recursos corticales de memoria desde un conjunto de patrones hasta otro, simplemente devolvería los antiguos reconocedores de patrones a la memoria y entonces realizaría la nueva asignación. Este tipo de "recogida de basura" y de reasignación de memoria es una característica estándar de la arquitectura de muchos sistemas de software. En nuestro cerebro digital también haríamos copias de seguridad de los recuerdos antes de desecharlos del neocórtex activo, una precaución que no podemos tomar en nuestros cerebros biológicos.

Existen varias técnicas matemáticas que pueden emplearse para implementar esta estrategia en el jerárquico reconocimiento de patrones autoorganizativo. El método que yo utilizaría sería el de los modelos ocultos jerárquicos de Márkov; por varias razones. Desde un punto de vista personal, llevo varias décadas familiarizado con este método, ya que lo he venido utilizando desde los primeros sistemas de reconocimiento del habla y del lenguaje natural en la década de 1980. Desde la perspectiva de la disciplina en general, existe una mayor experiencia acumulada con los modelos de Márkov que con cualquier otra estrategia para el reconocimiento de patrones. También son ampliamente utilizados en la comprensión del lenguaje natural. Muchos sistemas NLU utilizan técnicas que, por lo menos matemáticamente, son similares a HHMM.

Téngase en cuenta que todos los sistemas según el modelo oculto de Márkov son completamente jerárquicos. Algunos solo admiten unos pocos niveles de jerarquía; por ejemplo, pasar de estados acústicos a fonemas y de ahí a palabras. Para

construir un cerebro, sería deseable permitir que nuestro sistema creara tantos nuevos niveles de jerarquía como fuera necesario. Además, la mayoría de los sistemas según el modelo oculto de Márkov no son completamente autoorganizativos. Algunos tienen conexiones fijas, aunque estos sistemas podan de forma efectiva muchas de sus conexiones primigenias al no permitirlas desarrollar ningún peso asociado a sus conexiones. En las décadas de 1980 y de 1990, nuestros sistemas podaban automáticamente las conexiones cuyos pesos se encontraban por debajo de un cierto nivel y también permitían la creación de nuevas conexiones para modelizar mejor los datos de entrenamiento y para aprender sobre la marcha. Soy de la opinión de que un requisito fundamental es el de permitir al sistema crear de forma flexible sus propias topologías basándose en los patrones a los que está expuesto durante el aprendizaje. La técnica matemática de programación lineal la podemos usar para asignar óptimamente conexiones a los nuevos reconocedores de patrones.

Nuestro cerebro digital también soportará un importante grado de redundancia para cada patrón, especialmente para aquellos que aparecen frecuentemente. Esto permite un sólido reconocimiento de los patrones habituales y también es uno de los métodos fundamentales en la consecución del reconocimiento invariable de las diferentes formas de un patrón. Sin embargo, necesitaremos reglas que delimiten la cantidad de redundancia permitida, ya que no es deseable utilizar cantidades excesivas de memoria en patrones de bajo nivel que sean muy habituales.

Las reglas con respecto a la redundancia, a los umbrales de reconocimiento y al efecto sobre el umbral de una indicación del tipo “este patrón es esperado” suponen unos pocos ejemplos de parámetros generales fundamentales que afectan el rendimiento de este tipo de sistemas autoorganizativos. En un principio, yo fijaría estos parámetros atendiendo a mi intuición, pero luego serían optimizados mediante un algoritmo genético.

Un tema muy importante es la educación del cerebro, ya sea este biológico o esté basado en software. Tal y como he expuesto anteriormente, un sistema jerárquico de reconocimiento de patrones (digital o biológico) solo aprenderá unos dos, o preferiblemente uno, niveles jerárquicos a la vez. Para arrancar el sistema yo empezaría con redes jerárquicas previamente entrenadas que ya hayan aprendido sus lecciones de reconocimiento del habla humana, de caracteres impresos y de estructuras del lenguaje natural. Un sistema así sería capaz de leer documentos en lenguaje natural, pero solo podría dominar aproximadamente un nivel conceptual a la vez. Los niveles previamente aprendidos proporcionarían una base relativamente estable para aprender el siguiente nivel. El sistema podría leer los mismos documentos una y otra vez, consiguiendo nuevos niveles conceptuales con cada nueva lectura. De forma similar, las personas releen y logran una comprensión más profunda de los textos. En la web están disponibles miles de millones de

documentos. La propia Wikipedia, en su versión en inglés, contiene unos cuatro millones de artículos.

Yo también establecería un módulo de pensamiento crítico que realizaría un continuo escaneo de fondo de todos los patrones existentes y que revisaría su compatibilidad con los otros patrones (ideas) de este neocórtex basado en software. En nuestros cerebros biológicos no tenemos esta capacidad, razón por la cual las personas pueden mantener con ecuanimidad pensamientos completamente inconsistentes. En cuanto se identificara una idea inconsistente, el módulo digital se lanzaría a la búsqueda de una solución que incluiría sus propias estructuras corticales, así como toda la extensa literatura existente al respecto. Una solución puede limitarse a la determinación de que una de las ideas inconsistentes es simplemente incorrecta, ya que entra en contradicción con la evidencia que muestran los datos. Una solución más elaborada sería encontrar una idea perteneciente a un nivel conceptual más alto que resolviera la aparente contradicción proporcionando una perspectiva que explicara dicha idea. El sistema procedería a incorporar esta solución como un nuevo patrón y lo enlazaría con las ideas que inicialmente provocaron la búsqueda de la solución. Este módulo de pensamiento crítico se estaría ejecutando permanentemente a modo de tarea en segundo plano. Si el cerebro humano hiciera lo mismo, se trataría de algo muy beneficioso.

Personalmente, también proporcionaría un módulo que identificara cuestiones sin zanjar en todas las disciplinas. A modo de tarea permanente en segundo plano, buscaría soluciones a dichas cuestiones en otras áreas de conocimiento totalmente diferentes. Como ya he señalado, el conocimiento del neocórtex consiste en patrones de patrones profundamente asentados y por tanto es completamente metafórico. Podemos usar un patrón para encontrar una solución o posible estrategia en otro campo aparentemente inconexo.

Recuérdese por ejemplo la metáfora que utilicé en el capítulo 4 que ponía en relación los movimientos aleatorios de las moléculas de un gas y los movimientos aleatorios del cambio evolutivo. Las moléculas de un gas se mueven aleatoriamente sin ninguna dirección aparente. Pese a ello, literalmente todas las moléculas del gas contenidas en un matraz, después del tiempo necesario, acaban por abandonar el matraz. Ya señalé que esto nos orienta sobre una cuestión importante que tiene que ver con la evolución de la inteligencia. Al igual que las moléculas del gas, los cambios evolutivos también se producen en todas las direcciones y sin dirección aparente. Sin embargo, podemos observar un desplazamiento hacia una mayor complejidad y una mayor inteligencia; muestra de ello es el logro supremo de la evolución: el desarrollo de un neocórtex capaz de realizar pensamiento jerárquico. Así, observando un campo diferente (la termodinámica) podemos comprender cómo procesos aparentemente faltos de objetivo y sin dirección pueden dar lugar a un resultado que sí que tiene sentido en otro campo (el de la evolución biológica).

Ya he hecho referencia anteriormente a la perspectiva utilizada por Charles Lyell para explicar los cambios de las formaciones rocosas mediante pequeñas corrientes de agua que con el tiempo producen grandes valles, y cómo esta perspectiva inspiró a Charles Darwin para realizar una observación similar sobre los pequeños cambios en las características de diferentes organismos que pertenecen a una misma especie. Esta búsqueda de metáforas constituiría otro proceso permanente en segundo plano.

Deberíamos poner los medios para poder repasar simultáneamente múltiples listas, ya que esto equivale a permitir pensamiento estructurado. Una lista podría ser la relación de condiciones que debe satisfacer la solución de un problema. Cada paso puede dar lugar a una búsqueda recursiva a través de la jerarquía de ideas ya existente o a una búsqueda a través de la literatura disponible. Aparentemente, el cerebro humano solo es capaz de manejar cuatro listas simultáneamente (si no cuenta con la ayuda de herramientas tales como los ordenadores). Sin embargo, no hay razón para que un neocórtex artificial sufra la misma limitación.

También será deseable mejorar nuestros cerebros artificiales con el tipo de inteligencia que los ordenadores siempre han dominado, es decir, la capacidad para dominar con precisión enormes bases de datos y de implementar rápida y eficientemente los algoritmos conocidos. Wolfram Alpha combina de forma extraordinaria una gran cantidad de métodos científicos conocidos y los utiliza para recolectar datos cuidadosamente. Este tipo de sistemas va a continuar mejorando, ya que el Dr. Wolfram ha detectado un descenso exponencial en los índices de errores.

Por último, nuestro nuevo cerebro necesita tener un propósito. Un propósito viene definido por una serie de objetivos. En el caso de nuestros cerebros biológicos, nuestros objetivos vienen definidos por los centros de placer y de miedo que hemos heredado con nuestro cerebro antiguo. Inicialmente, estas tendencias primitivas fueron fijadas por la evolución biológica para favorecer la supervivencia de la especie. Sin embargo, el neocórtex nos ha permitido sublimarlas. El objetivo de Watson era responder a las preguntas de *Jeopardy!* Otro objetivo fácilmente definible podría ser pasar el test de Turing. Para ello, un cerebro digital necesitaría de una narración humana sobre la ficción de su propia historia, de manera que pudiera aparentar ser un humano biológico. También tendría que hacerse el tonto de forma considerable, ya que cualquier sistema que mostrara un conocimiento como por ejemplo el de Watson sería rápidamente desenmascarado.

Y lo que es más interesante todavía, a nuestro nuevo cerebro le podríamos dotar de un objetivo más ambicioso, como por ejemplo contribuir a mejorar el mundo. Por supuesto, un objetivo de estas características hace emerger muchas preguntas: ¿mejor para quién? ¿de qué manera mejor? ¿para los humanos biológicos? ¿para todos los seres conscientes? En ese caso, ¿quién o qué es consciente?

A medida que los cerebros no biológicos se vuelvan tan capaces como los

biológicos a la hora de infringir cambios en el mundo, y en último término sean muchos más capaces de ello que los cerebros biológicos no mejorados, tendremos que plantearnos su educación moral. Una buena forma de empezar sería una antigua idea procedente de nuestras tradiciones religiosas: la regla de oro.

* “Patch-clamp robot“ en el original.
* “Feedforward neural net“ en el original.
* Estas tres palabras las hemos dejado sin traducir porque obviamente tienen que ser pronunciadas en inglés para que el experimento descrito tenga sentido.
* Véase la nota anterior.
* Siglas en inglés que corresponden a “hidden Markov models”.
* “Skunk works project” en el original.
* Aquí tampoco hemos traducido las dos palabras, “one” (uno) y “two” (dos).
* Literalmente, oveja eléctrica.
* “Overfitting” en el original.
* “Recursive cortical network” en el original.
* Aquí hay que entender la dinámica del juego. En *Jeopardy!* no se da la respuesta tal cual, sino que hay que integrarla dentro de una pregunta del tipo ¿Qué es ?
* Novela escrita por Douglas Adams en 1979.
* “Tail of a language” en el original.
* “Expert manager” en el original.
** “Orthogonal” en el original.
* Novela de Charles Dickens publicada en 1859.

CAPÍTULO OCHO

# La mente como ordenador

Con una forma un tanto similar a una rebanada de pan rústico francés, nuestro cerebro es un laboratorio químico muy concurrido que bulle con interminables conversaciones neuronales. Imagínese así el cerebro: ese brillante collado del ser, ese parlamento de células color gris como los ratones, esa factoría de sueños, ese pequeño tirano dentro de una bola de huesos, ese corrillo de neuronas que canta todas las jugadas, esa pequeña omnipresencia, ese caprichoso templo de placer, ese arrugado armario cuyas baldas pueblan el cráneo como lo hacen las ropas de una bolsa de deporte demasiado llena.

—DIANE ACKERMAN

Los cerebros existen porque la distribución de recursos necesaria para la supervivencia y los peligros que amenazan la supervivencia varían según el espacio y el tiempo.

—JOHN M. ALLMAN

La geografía moderna del cerebro le da un delicioso toque anticuado como pasa con un mapa medieval donde el mundo conocido aparece rodeado por *terra incognita* plagada de monstruos.

—DAVID BAINBRIDGE

En matemáticas las cosas no se comprenden. Uno simplemente se acostumbra a ellas.

—JOHN VON NEUMANN

Desde que a mediados del siglo XX apareció el ordenador, no se ha dejado de debatir, no solo sobre sus capacidades últimas, sino sobre si el propio cerebro humano podría ser considerado una forma de ordenador. En lo que respecta a esta última pregunta, el consenso ha oscilado entre la visión que sostiene que estos dos tipos de entidades procesadoras de información son esencialmente iguales y la visión que sostiene que son fundamentalmente diferentes. De modo que, *¿es* el cerebro un ordenador?

Cuando los ordenadores se volvieron una cuestión popular en la década de 1940, inmediatamente fueron considerados máquinas pensantes. El ENIAC, introducido en 1946, fue descrito por la prensa como un “cerebro gigante”. A medida que los ordenadores se volvieron asequibles comercialmente durante la siguiente década, la publicidad solía referirse a ellos como cerebros capaces de realizar tareas que los cerebros biológicos normales no podían igualar.

**Anuncio de 1957 que muestra la popular idea de que el ordenador es un cerebro gigante.**

Rápidamente, los programas de ordenador permitieron que las máquinas cumplieran con las expectativas puestas en ellos. El "solucionador universal de problemas"* creado en 1959 por Herbert A. Simon, J. C. Shaw y Allen Newell en la Carnegie Mellon University fue capaz de demostrar un teorema que los matemáticos Bertrand Russell (1872–1970) y Alfred North Whitehead (1861–1947) no habían podido resolver en su famosa obra de 1913 llamada *Principia Mathematica*. Lo que décadas posteriores dejaron claro fue que los ordenadores podían fácilmente superar con mucho las capacidades humanas en ejercicios intelectuales tales como la resolución de problemas matemáticos, el diagnóstico de enfermedades y el ajedrez, pero que sin embargo tenían dificultades a la hora de hacer que un robot se atara los cordones de los zapatos o para comprender el lenguaje de sentido común que un niño

de 5 años es capaz de entender. Solo ahora los ordenadores están empezando a dominar este tipo de capacidades. Irónicamente, la evolución de la inteligencia de los ordenadores ha transcurrido en dirección opuesta a la del desarrollo humano.

La cuestión de si el ordenador y el cerebro humano se encuentran a un mismo nivel equivalente sigue siendo a día de hoy una cuestión controvertida. En la introducción ya mencioné que había millones de enlaces que mencionaban la complejidad del cerebro humano. De forma similar, una búsqueda mediante Google de "citas: el cerebro no es un ordenador" también da como resultado millones de enlaces. En mi opinión, declaraciones como estas son como decir "la salsa de manzana no es manzana". Técnicamente dicha declaración es verdad, pero se puede hacer salsa de manzana partiendo de una manzana. Quizá fuera más apropiada la aserción que dice "los ordenadores no son procesadores de texto". Es cierto que un ordenador y un procesador de texto se encuentran en diferentes niveles conceptuales; sin embargo, un ordenador puede convertirse en un procesador de texto si ejecuta un software de procesamiento de textos y lo contrario no es posible. De forma similar, un ordenador puede convertirse en un cerebro si ejecuta software cerebral. Esto es lo que investigadores entre los que me incluyo estamos intentando hacer.

Por tanto, la cuestión es si podremos o no encontrar un algoritmo que convierta un ordenador en una entidad que sea equivalente a un cerebro humano. Después de todo, y debido a su universalidad innata (dependiente solo de su capacidad), un ordenador puede ejecutar cualquier algoritmo. Por otra parte, el cerebro humano ejecuta un conjunto de algoritmos específicos. Sus métodos son inteligentes en el sentido de que permite un importante grado de plasticidad y de reestructuración de sus propias conexiones basándose en su experiencia. No obstante, estas funciones son susceptibles de ser emuladas mediante software.

La universalidad de la computación (la idea de que un ordenador de propósito general puede implementar cualquier algoritmo) y el poder de esta idea emergieron al mismo tiempo que las primeras máquinas. Existen cuatro conceptos fundamentales que subyacen tras la universalidad y viabilidad de la computación y su aplicabilidad a nuestro pensamiento. Merece la pena que los repasemos, ya que el propio cerebro hace uso de ellos. El primer concepto es la capacidad de comunicar, recordar y computar información de forma fiable. Hacia 1940, si se usaba la palabra "ordenador", la gente asumía se estaba hablando de un ordenador analógico en el que los números venían representados por diferentes niveles de voltaje y en el que los componentes especializados podían realizar operaciones aritméticas tales como sumas y multiplicaciones. Sin embargo, una gran limitación de los ordenadores analógicos era que estaban plagados de fallos de precisión. Los números solo podían ser representados con una precisión de más o menos una centésima, y como los niveles de voltaje que los representaban eran procesados mediante números crecientes de operadores aritméticos, los errores se acumulaban. Si lo que se

pretendía era realizar más de un puñado de cálculos, los resultados se volvían tan imprecisos que acababan por perder su significado.

Cualquiera que pueda recordar la época en la que se grababa música mediante magnetófonos analógicos se acordará de este efecto. Se producía una notable degradación en la primera copia, que era un poco más ruidosa que el original. (Recuérdese que "el ruido" representa imprecisiones aleatorias). La copia de la copia era todavía más ruidosa y cuando se llegaba a la décima generación la copia era casi completamente ruido. Se daba por asumido que el mismo problema también plagaría el emergente mundo de los ordenadores digitales. Esta preocupación es entendible si se considera la comunicación de información digital a través de un canal, ya que no hay ningún canal perfecto, todos poseen un inherente índice de error. Supongamos que tenemos un canal que tiene una probabilidad de 0,9 de transmitir correctamente cada uno de los bits. Si envío un mensaje de un bit de longitud, la probabilidad de transmitirlo correctamente a través de dicho canal será de 0,9. Supongamos que envío dos bits Entonces la precisión es de $0{,}9^2 = 0{,}81$. ¿Qué pasa si envío un byte (ocho bits)? Entonces tengo menos de la mitad de posibilidades (0,43 para ser exactos) de enviarlo correctamente. La probabilidad de enviar de forma precisa cinco bytes es de más o menos el 1%.

Una solución obvia para evitar este problema es hacer el canal más preciso. Supongamos que el canal solo comete un error en un millón de bits. Si envío un fichero que consta de medio millón de bytes (más o menos el tamaño de un programa o base de datos simple), la probabilidad de transmitirlo correctamente es de menos del 2%, pese al alto grado de precisión inherente al canal. Dado que un error en un solo bit puede invalidar completamente un programa informático y otras formas de datos digitales, esta no es una situación satisfactoria. Independientemente de la precisión del canal, como la probabilidad de un error en la transmisión crece rápidamente con el tamaño del mensaje, podría parecer que nos encontráramos ante una barrera infranqueable.

Los ordenadores analógicos abordaban este problema por medio de una degradación elegante (en el sentido en que los usuarios solo planteaban problemas en los que se podían tolerar pequeños errores). Sin embargo, si los usuarios de ordenadores analógicos se limitaban a un conjunto restringido de cálculos, los ordenadores sí que eran útiles en cierta medida. Por otra parte, los ordenadores digitales necesitan de una comunicación permanente, no solo desde un ordenador a otro, sino en el interior del propio ordenador. Existe comunicación desde su memoria hasta la unidad central de procesamiento y viceversa. En el interior de la unidad central de procesamiento se produce comunicación desde un registro a otro y de ida y vuelta a la unidad aritmética. Incluso en el interior de la unidad aritmética se produce comunicación desde un registro de bits y otro. La comunicación se produce

por todos los sitios y a todos los niveles. Si tenemos en cuenta que los índices de error aumentan rápidamente con el aumento de la comunicación y que un error en un solo bit puede destruir un proceso íntegramente, la comunicación digital estaba condenada, o al menos eso es lo que parecía entonces.

Curiosamente, esta era la opinión generalizada hasta que apareció en escena el matemático norteamericano Claude Shannon (1916–2001) y demostró cómo se puede producir comunicación precisa de forma arbitraria usando incluso los canales de comunicación menos fiables. Lo que Shannon sostuvo en su histórico trabajo "A Mathematical Theory of Communication", publicado en el *Bell System Technical Journal* en julio y octubre de 1948, sobre el teorema de la codificación de canales con ruido*, fue que si se tenía acceso a un canal con un índice de error cualquiera (excepto un índice de error de exactamente el 50% por bit, ya que eso significaría que el canal solo estaría transmitiendo puro ruido), se puede transmitir un mensaje en el cual el índice de error es tan pequeño como se desee. En otras palabras, el índice de error de la transmisión puede ser de un bit por cada $n$ bits, siendo $n$ tan grande como lo definamos. De manera que si, por ejemplo, tenemos un caso extremo en el que un canal transmite correctamente bits de información solo el 51% de las veces (es decir, que transmite el bit correcto solo un poquito más a menudo que el bit incorrecto), entonces se pueden transmitir mensajes de manera que solo un bit de entre un millón sea incorrecto, o un bit de entre un billón, o de entre un billón de billones.

¿Cómo puede ser esto posible? Por la redundancia. Puede que hoy esto parezca obvio, pero en aquel momento no lo era. Pongamos un ejemplo sencillo. Si transmito cada bit tres veces y me quedo con el resultado que más se repite, habré aumentado sustancialmente la fiabilidad del resultado. Si este no es lo suficientemente bueno, solo se ha de aumentar la redundancia hasta que se alcance la fiabilidad que se necesite. La simple repetición de la información es la manera más sencilla de alcanzar índices de precisión tan altos como se desee a partir de canales de baja precisión. Sin embargo, esta estrategia no es la más eficiente. El trabajo de Shannon, que puso las bases de la teoría de la información, presentaba métodos óptimos de detección de errores y códigos de corrección que pueden alcanzar *cualquier* precisión deseada a través de *cualquier* canal no aleatorio.

Los lectores más mayores se acordarán de los módems telefónicos que transmitían información a través de ruidosas líneas de teléfono analógicas. Estas líneas producían silbidos y ruidos como de pequeñas explosiones claramente audibles y muchas otras formas de distorsión, y sin embargo eran capaces de transmitir datos digitales con índices de precisión muy elevados gracias al teorema de los canales con ruido de Shannon. La misma cuestión y la misma solución existe en el caso de la memoria digital. ¿Se ha preguntado alguna vez cómo CDs, DVDs y discos programables siguen ofreciendo resultados fiables incluso después de que el

disco se haya caído al suelo y se haya arañado? De nuevo, hay que agradecérselo a Shannon.

La computación consta de tres elementos: comunicación (que, como ya he dicho, se produce tanto en el interior de los ordenadores como entre ellos), memoria y puertas lógicas (que realizan las funciones aritméticas y lógicas). La precisión de las puertas lógicas también puede hacerse que sea todo lo alta que se desee mediante detección de errores y códigos de corrección similares. Es gracias al teorema y la teoría de Shannon por lo que podemos manejar algoritmos y datos digitales arbitrariamente grandes y complejos sin que los errores desvirtúen o destruyan los procesamientos. Es importante reseñar que el cerebro también utiliza el principio de Shannon, ¡aunque la evolución del cerebro humano sea claramente muy anterior al propio Shannon! La mayoría de los patrones o ideas (una idea también es un patrón), como ya hemos visto, están almacenados en el cerebro con una notable cantidad de redundancia. Una de las razones principales de la redundancia del cerebro es la inherente falta de fiabilidad de los circuitos neuronales.

La segunda idea importante en la que descansa la era de la información es la que ya mencioné anteriormente: la universalidad de la computación. En 1936, Alan Turing describió su "máquina de Turing", que en realidad no era una máquina sino otro experimento mental. Su teórico ordenador consiste en una cinta de memoria infinitamente larga con un 1 o un 0 en cada cuadro. El input de la máquina es introducido en esta cinta, ya que la máquina puede leer un cuadro cada vez. La máquina también contiene una tabla de reglas, lo que viene a ser un programa almacenado, que consta de estados numerados. Cada regla especifica una acción si el cuadro que está siendo leído es un 0 y otra acción diferente si el cuadro leído es un 1. Las posibles acciones a realizar incluyen escribir un 0 o un 1 en la cinta, mover la cinta un cuadro hacia la derecha o hacia la izquierda, o detenerse. Así, cada estado especifica el número del siguiente estado en el que va a entrar la máquina.

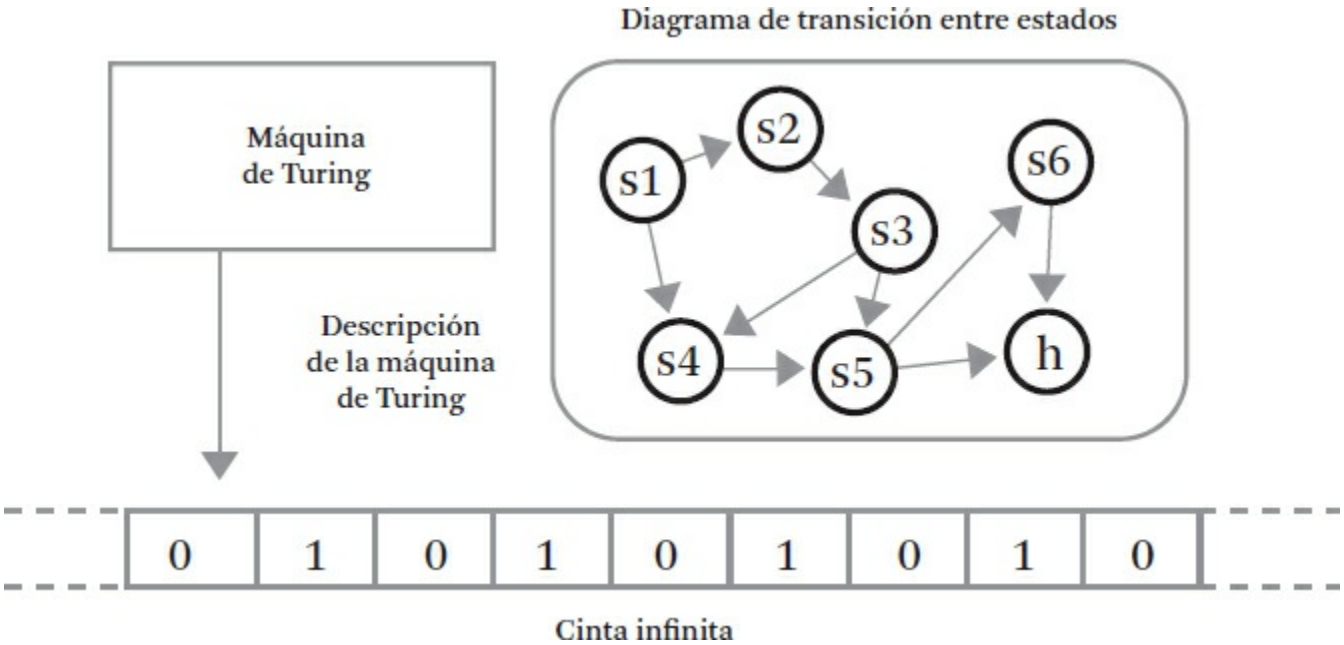
El input de la máquina de Turing es introducido en la cinta. El programa se ejecuta y cuando la máquina se detiene es que ha completado su algoritmo y el output del proceso permanece en la cinta. Nótese que aunque la cinta tenga una longitud teóricamente infinita, cualquier programa que no caiga en un bucle infinito solo utilizará una porción finita de la cinta, de manera que si nos limitamos a usar una cinta finita, la máquina continuará resolviendo un conjunto útil de problemas.

Si la máquina de Turing parece simple es porque ese fue el objetivo de su inventor. Turing quería que su máquina fuera tan simple como fuera posible (pero, parafraseando a Einstein, que no fuera más simple que lo que fuera posible). Turing y Alonzo Church (1903–1995), su antiguo profesor, desarrollaron la tesis Church-Turing, que sostiene que si un problema que pueda ser introducido en una máquina de Turing no puede ser resuelto por ella, es que no es resoluble por *ninguna* máquina

que esté sujeta a las leyes naturales. Aunque la máquina de Turing solo posee un puñado de comandos y solo procesa un bit cada vez, puede calcular cualquier cosa que cualquier ordenador pueda calcular. Otra manera de decirlo es que cualquier máquina que sea "completamente Turing" (es decir, que posea unas capacidades equivalentes a las de una máquina de Turing) puede calcular cualquier algoritmo (cualquier procedimiento que podamos definir).

Interpretaciones "fuertes" de la tesis Church-Turing sostienen una equivalencia fundamental entre lo que un humano puede pensar o conocer y lo que es calculable por una máquina. La idea básica es que el cerebro humano está asimismo sujeto a las leyes naturales, por lo que su capacidad para procesar información no puede superar a la de una máquina (y por tanto a la de una máquina de Turing).

Podemos otorgar a Turing el honor de establecer las bases teóricas de la computación mediante su trabajo de investigación de 1936. Sin embargo, es importante tener en cuenta que Turing estaba profundamente influenciado por la conferencia que un matemático húngaro-norteamericano llamado John von Neumann (1903–1957) dio en Cambridge en el año 1935 sobre su idea de programa almacenado, una idea consagrada en la máquina de Turing (1). A su vez, von Neumann estaba influenciado por el trabajo de investigación de Turing del año 1936, que de forma elegante estableció los principios de la computación y se convirtió en lectura obligada para sus colegas de finales de la década de 1930 y principios de la de 1940 (2).




**Un diagrama de bloques de una máquina Turing con un cabezal que lee y escribe la cinta y un programa interno consistente en estados en transición.**

En dicho trabajo, Turing anuncia otro descubrimiento inesperado, el de los problemas irresolubles. Se trata de problemas bien definidos con soluciones únicas cuya existencia puede ser probada. Sin embargo, también se puede probar que nunca

podrán ser calculadas por una máquina de Turing, que es lo mismo que decir por *cualquier* máquina, lo cual supuso el final del dogma decimonónico que sostenía que los problemas bien definidos podían en último término ser resueltos. Turing demostró que existen tantos problemas irresolubles como solubles. El matemático y filósofo austro-norteamericano Kurt Gödel llegó a una conclusión similar en el año 1931 mediante su "teorema de incompletud". Así, nos encontramos en la paradójica situación de poder definir un problema, de probar que tiene una solución única existente, y sin embargo de saber que dicha solución nunca podrá ser encontrada.

Turing demostró que en esencia la computación se basa en un mecanismo muy simple. Como la máquina de Turing (y por tanto cualquier ordenador) es capaz de basar su curso de acción futuro en los resultados que ya ha computado, también es capaz de tomar decisiones y de modelizar arbitrariamente complejas jerarquías de información.

En 1939 Turing diseñó una calculadora electrónica llamada Bombe que ayudó a decodificar mensajes encriptados por la máquina nazi de codificación llamada Enigma. En 1943 un equipo de ingenieros influenciado por Turing completó lo que podría ser considerado como el primer ordenador, el Colossus, el cual permitió que los aliados siguieran decodificando mensajes procedentes de versiones más sofisticadas de Enigma. Bombe y Colossus fueron diseñados para una única tarea y no podían ser reprogramados para ninguna otra. Sin embargo, realizaban esta tarea brillantemente y tienen el mérito de haber permitido que los aliados superaran la ventaja de tres a uno que la *Luftwaffe* alemana disfrutaba sobre la *Royal Air Force* británica y que ganaran la crucial Batalla de Inglaterra, así como que los aliados siguieran anticipándose a las tácticas nazis durante toda la guerra.

Sobre estos cimientos, John von Neumann creó la arquitectura del ordenador moderno, que es lo que representa nuestra tercera gran idea. Bajo el nombre de la máquina von Neumann sigue siendo la estructura nuclear de básicamente todos los ordenadores construidos durante los últimos sesentaisiete años, desde el microcontrolador de su lavadora hasta el más grande de los superordenadores. En un trabajo de investigación que data del 30 de junio de 1945 y que lleva el título de "First Draft of a Report on the EDVAC", von Neumann presentó las ideas que han dominado la computación desde entonces (3). El modelo de von Neumann incluye una unidad central de procesamiento donde las operaciones aritméticas y lógicas son llevadas a cabo, una unidad de memoria donde se almacenan el programa y los datos, un almacenamiento masivo, un contador de programa y los canales de input/output. Aunque este trabajo tenía la intención de ser un documento para un proyecto interno, se ha convertido en la biblia de los diseñadores de ordenadores. Nunca se sabe cuando un memorándum interno aparentemente rutinario acabará por revolucionar el mundo.

La máquina de Turing no se diseñó para ser práctica. Los teoremas de Turing no prestaban atención a la eficiencia a la hora resolver problemas, sino al examen del rango de problemas que en teoría podrían ser resueltos mediante computación. El objetivo de von Neumann, por otra parte, era el de crear el concepto factible de una máquina computacional. Su modelo remplaza al de computaciones de un bit creado por Turing por otro de palabras de múltiples bits (por lo general, de algún múltiplo de ocho bits). La cinta de memoria de Turing es secuencial, de manera que los programas de la máquina Turing se pasan una cantidad de tiempo excesiva moviendo la cinta hacia adelante y hacia atrás para almacenar y recuperar resultados intermedios. Por el contrario, la memoria de von Neumann es de acceso aleatorio, por lo que cualquier ítem de datos puede ser recuperado inmediatamente.

Una de las ideas fundamentales de von Neumann es el programa almacenado, introducido por él una década antes y que dota al programa del mismo tipo de acceso aleatorio a la memoria del que disfrutan los datos (y que además suele encontrarse en el mismo bloque de memoria). Esto permite que el ordenador sea reprogramado para diferentes tareas, así como para la automodificación del código si el programa de almacenamiento es de escritura, lo cual permite que se dé una poderosa forma de recursión. Hasta ese momento, literalmente todos los ordenadores, incluido el Colossus, estaban fabricados para una tarea específica. El programa almacenado posibilita que un ordenador sea verdaderamente universal, y por tanto que cumpla con la idea de Turing sobre la universalidad de la computación.

Otro aspecto fundamental de la máquina von Neumann es que todas las instrucciones incluyen un código de operación que especifica la operación aritmética o lógica a realizar, así como la dirección de un operando de la memoria.

La idea de von Neumann sobre cómo debía ser la arquitectura de un ordenador fue presentada mediante su publicación sobre el diseño del EDVAC, un proyecto que dirigió junto a sus colaboradores J. Presper Eckert y John Mauchly. El propio EDVAC no empezó a funcionar hasta 1951, cuando ya existían otros ordenadores de programa almacenado tales como el Manchester Small-Scale Experimental Machine, ENIAC, EDSAC y BINAC, todos ellos profundamente influenciados por el trabajo de von Neumann y en los que Eckert y Mauchly habían trabajado como diseñadores. Von Neumann hizo contribuciones directas al diseño de varias de estas máquinas, incluyendo la última versión de ENIAC, que ya daba soporte a un programa almacenado.

La arquitectura de von Neumann tuvo algunos precursoras, aunque, con la sorprendente excepción de una de ellas, ninguna se corresponde con una verdadera máquina von Neumann. En 1944, Howard Aiken presentó el Mark I, el cual poseía un elemento programable pero que no utilizaba un programa almacenado. Leía instrucciones procedentes de una cinta de papel perforado y luego ejecutaba cada

comando inmediatamente. También le faltaba una instrucción de la rama condicional.

En 1941, el científico alemán Konrad Zuse (1910–1995) creó el ordenador Z-3. También leía su programa a partir de una cinta (que en este caso venía codificada en una película) y tampoco poseía una instrucción de rama condicional. Curiosamente, Zuse estaba financiado por el Centro de Investigaciones Aéreas de Alemania, que utilizó el dispositivo para estudiar las vibraciones de las alas. Sin embargo, su petición de financiación al Gobierno nazi para remplazar sus relés con tubos de vacío fue rechazada. Los nazis juzgaban a la computación como algo "no importante para la guerra". En mi opinión, esta forma de pensar tuvo profundas implicaciones en el resultado de la contienda.

De hecho, sí que existe un genuino precursor de la idea de von Neumann ¡que proviene de un siglo anterior! La máquina analítica del inventor y matemático inglés Charles Babbage, cuya primera descripción data de 1837, incorporaba las ideas de von Neumann y contenía un programa almacenado por medio de cartones perforados que había tomado prestados del telar de Jacquard (4). Su memoria de acceso aleatorio incluía 1 000 palabras de 50 dígitos decimales cada una (el equivalente a unos 21 kilobytes). Cada instrucción incluía un código de operación y un número de operando, igual que los lenguajes máquina modernos. Incluía ramas y bucles condicionales, de manera que era una verdadera máquina von Neumann. Estaba basada por completo en marchas mecánicas y parece ser que Analytical Engine superaba las capacidades de diseño y de organización de Babbage. Él construyó partes, pero la máquina nunca funcionó. No está claro si los pioneros de los ordenadores del siglo XX, incluido von Neumann, estaban al corriente de la obra de Babbage.

El ordenador de Babbage desembocó en la creación del campo de la programación de software. La escritora inglesa Ada Byron (1815–1852), Condesa de Lovelace y única hija legítima del poeta Lord Byron, fue la primera programadora informática del mundo. Escribió programas para Analytical Engine que, como el ordenador nunca funcionó, tenía que depurar en su propia cabeza, una práctica bien conocida para los programadores de hoy en día que se llama "verificación de tablas". Tradujo un artículo que el matemático italiano Luigi Menabrea había escrito sobre Analytical Engine y le añadió amplias notas de su puño y letra. Así, escribió: "Analytical Engine teje patrones algebraicos, tal y como el telar de Jacquard teje flores y hojas". Además, proporcionó las que quizá son las primeras especulaciones sobre la viabilidad de la inteligencia artificial. Sin embargo, llegó a la conclusión de que Analytical Engine "no pretendía dar origen a nada".

El invento de Babbage es milagroso si se considera la época en la que vivió y trabajó. Sin embargo, a mediados del siglo XX sus ideas se habían perdido en las nieblas del pasado (aunque después fueron redescubiertas). Fue von Neumann quien conceptualizó y articuló los principios fundamentales de los ordenadores tal y como

los conocemos hoy, y el mundo se lo reconoce citando continuamente la máquina von Neumann como el modelo principal de la computación. Sin embargo, tenga en cuenta que la máquina von Neumann transmite datos continuamente entre sus diversas unidades y en el interior de estas, de manera que no pudo ser construida sin los teoremas de Shannon y los métodos que él ideó para la transmisión y almacenamiento de información digital fiable.

Esto nos lleva hasta la cuarta idea importante, que supone ir más allá de la conclusión de Ada Byron que sostenía que un ordenador no podía pensar creativamente ni encontrar los algoritmos fundamentales empleados por el cerebro para luego utilizarlos en convertir un ordenador en un cerebro. Este objetivo fue presentado por Alan Turing en su trabajo de investigación de 1950 "Computing Machinery and Intelligence", que incluye su famoso test de Turing para comprobar si una IA ha alcanzado un nivel de inteligencia humano.

En 1956, von Neumann empezó a preparar una serie de conferencias para la prestigiosa serie de conferencias Silliman en la Universidad de Yale. Debido a los estragos del cáncer, nunca llegó a dar dichas charlas ni llegó a completar el manuscrito a partir del cual tenían que organizarse. No obstante, este documento inacabado sigue siendo un brillante y profético presagio de lo que yo considero como el proyecto más grande e importante de la humanidad. Se publicó póstumamente en 1958 bajo el nombre *The Computer and the Brain*. Resulta razonable que el último trabajo de uno de los matemáticos más brillantes del último siglo y uno de los pioneros de la era de los ordenadores versara sobre la propia inteligencia. Este proyecto fue la primera investigación seria sobre el cerebro humano desde la perspectiva de un matemático y científico de la computación. Antes de von Neumann los campos de la ciencia de la computación y de la neurociencia eran dos islas sin nexo de unión.

Von Neumann comienza su exposición reflejando las similitudes y diferencias entre un ordenador y el cerebro humano. Teniendo en cuenta cuándo escribió el manuscrito, este es sorprendentemente certero. Se dio cuenta de que el output de las neuronas era digital (un axón o bien se dispara, o bien no se dispara). En aquella época, esto estaba muy lejos de ser algo obvio, ya que el output podría haberse tratado de una señal analógica. Por su parte, el procesamiento de las dendritas que desemboca en la neurona y en el cuerpo celular de la neurona es analógico, y describió sus cálculos como la suma de pesos de los inputs según un umbral. Este modelo sobre cómo funcionan las neuronas desembocó en el campo del conexionismo, el cual construye sistemas basándose en este modelo de neurona, tanto en lo que respecta al hardware como en lo que respecta al software. (Tal y como describí en el capítulo anterior, el primer sistema conexionista fue creado por Frank Rosenblatt a modo de un programa de software para un ordenador IBM 704 en

Cornell en 1957, inmediatamente después de que los esbozos de conferencias de von Neumann se hicieran públicos). Ahora disponemos de modelos más sofisticados sobre cómo las neuronas combinan inputs, pero la idea esencial del procesamiento analógico de los inputs de la dendrita mediante concentraciones de neurotransmisores sigue siendo válida.

Von Neumann hizo uso del concepto de la universalidad de la computación para llegar a la conclusión de que a pesar de que la arquitectura y los módulos componentes sean radicalmente diferentes en el caso del cerebro y del ordenador, podemos asegurar que una máquina von Neumann puede simular el procesamiento que realiza un cerebro. Sin embargo, el caso contrario no es cierto, ya que el cerebro no es una máquina von Neumann y no posee un programa almacenado como tal (aunque podemos simular una máquina de Turing muy simple en nuestras cabezas). Su algoritmo o métodos están implícitos es su estructura. Von Neumann llega a la correcta conclusión de que las neuronas pueden aprender patrones a partir de sus inputs (que a día de hoy está demostrado que están codificados parcialmente en la fuerza de las dendritas). Lo que en la época de von Neumann no se sabía es que el aprendizaje también tiene lugar a través de la creación y destrucción de conexiones entre neuronas.

Proféticamente, von Neumann señala que la velocidad del procesamiento neuronal es extremadamente lenta (del orden de cien cálculos por segundo), pero que el cerebro compensa esto a través del procesamiento masivamente paralelo, otra aportación fundamental y no obvia. Von Neumann sostenía que todas las $10^{10}$ neuronas del cerebro (un cálculo bastante aproximado, ya que a día de hoy se calcula que está entre las $10^{10}$ y las $10^{11}$ neuronas) procesaban al mismo tiempo. De hecho, todas las conexiones (que por neurona son una media de entre $10^3$ y $10^4$) realizan computación simultáneamente.

Las estimaciones y descripciones del procesamiento neuronal hechas por von Neumann son dignas de mención teniendo en cuenta el primitivo estado de la neurociencia en aquel tiempo. Sin embargo, un aspecto de su obra con el que estoy en desacuerdo es su opinión sobre la capacidad de memoria del cerebro. Él asume que el cerebro recuerda todos los inputs de por vida. Von Neumann asume una esperanza de vida media de 60 años, o lo que es lo mismo, unos $2 \cdot 10^9$ segundos. Con unos 14 inputs por segundo en cada neurona (que es un cálculo por lo menos tres órdenes de magnitud demasiado bajo) y contando con $10^{10}$ neuronas, obtiene una estimación de alrededor de $10^{20}$ bits para la capacidad de memoria del cerebro. Tal y como he señalado anteriormente, en realidad recordamos solo una fracción muy pequeña de nuestros pensamientos y experiencias, e incluso estos recuerdos no están almacenados a modo de patrones de bits a un nivel bajo (como ocurre con una imagen de video), sino como secuencias de patrones de más alto nivel.

A medida que von Neumann describe los mecanismos del cerebro, demuestra cómo un ordenador moderno podría proporcionar los mismos resultados pese a las evidentes diferencias existentes entre ambos. Los mecanismos analógicos del cerebro pueden ser simulados a través de mecanismos digitales, ya que la computación digital puede emular valores analógicos con un grado de precisión tan alto como se desee (y la precisión de la información analógica del cerebro es bastante baja). El masivo paralelismo del cerebro también puede ser simulado gracias a la gran ventaja en cuanto a velocidad que ofrecen los ordenadores de computación en serie (una ventaja que con el tiempo ha crecido enormemente). Además, también podemos hacer uso del procesamiento en paralelo de ordenadores mediante el uso en paralelo de máquinas von Neumann, que es exactamente la manera en la que funcionan los superordenadores de hoy en día.

Von Neumann llega a la conclusión de que los métodos del cerebro no pueden conllevar algoritmos secuenciales largos, ya que no hay más que observar lo rápido que los humanos son capaces de tomar decisiones y la tan lenta velocidad de computación de las neuronas. Cuando un jugador en la tercera base decide lanzar la pelota de béisbol a la primera en vez de a la segunda base, toma esta decisión en una fracción de segundo, que es solo el tiempo necesario para que las neuronas realicen un puñado de ciclos. Acertadamente, von Neumann llega a la conclusión de que las extraordinarias capacidades del cerebro son producto del procesamiento simultáneo de información que sus 100 000 millones de neuronas son capaces de realizar. Así, tal y como he señalado, el córtex visual realiza sofisticados juicios visuales en solo tres o cuatro ciclos neuronales.

El cerebro posee una considerable plasticidad, que es lo que nos permite aprender. Sin embargo, la plasticidad de un ordenador es mucho mayor, ya que cambiando su software se pueden reestructurar completamente sus métodos. Por tanto, un ordenador será capaz de emular al cerebro, pero no al revés.

Cuando von Neumann comparaba la capacidad de la organización masivamente paralela del cerebro con los escasos ordenadores de su época, era evidente que el cerebro poseía una memoria y velocidad muy superiores. A día de hoy ya se ha construido el primer superordenador cuyas especificaciones se corresponden con algunas de las estimaciones más conservadoras en cuanto a la velocidad necesaria para simular funcionalmente el cerebro humano (unas $10^{16}$ operaciones por segundo) (5). Yo calculo que este nivel de computación costará $1 000 a principios de la década de 2020. Nos encontramos todavía más cerca en cuanto a la capacidad de memoria. Aunque su manuscrito fue escrito sorprendentemente pronto en lo que respecta a la historia del ordenador, von Neumann confiaba en que tanto el hardware como el software de la inteligencia humana acabarían por hacer su aparición, razón por la cual había preparado las conferencias.

Von Neumann estaba enormemente afectado por el incremento en el ritmo del progreso y sus profundas implicaciones para el futuro de la humanidad. En 1957, un año después de su muerte, otro matemático llamado Stan Ulam le citó como el autor de la frase dicha a principios de la década de 1950 que reza: "el progreso cada vez más acelerado de la tecnología y los cambios en el modo de vida humano dan la apariencia de estar acercándose a algún tipo de singularidad esencial en la historia de nuestra raza más allá de la cual los asuntos humanos no podrán seguir siendo tal y como los conocemos". Se trata del primer uso conocido de la palabra "singularidad" en el contexto de la historia tecnológica humana.

La idea fundamental de von Neumann era que existe una equivalencia esencial entre un ordenador y un cerebro. Téngase en cuenta que la inteligencia emocional de los humanos biológicos forma parte de su inteligencia. Si la idea de von Neumann es correcta y si se acepta mi propio acto de fe en que una entidad no biológica que recree convincentemente la inteligencia (emocional y de otro tipo) de un humano biológico poseerá consciencia (véase el siguiente capítulo), entonces se tendría que concluir que existe una equivalencia esencial entre un ordenador *dotado del software adecuado* y una mente (consciente). De manera que ¿tiene razón von Neumann?

La mayoría de los ordenadores de hoy en día son completamente digitales, mientras que el cerebro humano combina métodos digitales y analógicos. Sin embargo, los métodos analógicos son fácil y rutinariamente recreados mediante métodos digitales con un nivel de precisión tan alto como se desee. El científico de la computación norteamericano Carver Mead (nacido en 1934) ha probado que podemos emular directamente los métodos analógicos del cerebro con silicio, cosa que ha demostrado mediante lo que él llama chips "neuromórficos" (6). Mead ha comprobado cómo esta estrategia puede ser miles de veces más eficiente que la emulación digital de los métodos analógicos. A medida que codifiquemos el algoritmo neocortical masivamente repetido, ganará sentido usar la estrategia de Mead. El Cognitive Computing Group de IBM, liderado por Dharmendra Modha, ha presentado chips que emulan a las neuronas y sus conexiones, incluyendo la capacidad de formar nuevas conexiones (7). Bajo el nombre de "SyNAPSE", uno de los chips proporcionó una simulación directa de 256 neuronas con alrededor de un cuarto de millón de conexiones sinápticas. El objetivo del proyecto es crear un neocórtex simulado con 10 000 millones de neuronas y 100 billones de conexiones (casi las de un cerebro humano) que utilice solo un kilovatio de potencia.

Tal y como von Neumann describió hace más de medio siglo, el cerebro es extremadamente lento pero masivamente paralelo. Los circuitos digitales de hoy en día son por lo menos 10 millones de veces más rápidos que los interruptores electroquímicos del cerebro. Por otra parte, todos y cada uno de los 300 millones de

reconocedores de patrones neocorticales del cerebro procesan simultáneamente, y todas y cada una de las miles de billones de conexiones interneuronales computan potencialmente al mismo tiempo. Sin embargo, la cuestión fundamental para conseguir el hardware necesario para modelizar con éxito un cerebro humano es la capacidad general de memoria y el rendimiento necesarios. Por tanto, no necesitamos copiar directamente la arquitectura del cerebro. Esta sería una estrategia muy ineficiente e inflexible.

Hagamos una estimación sobre cuáles son dichos requisitos de hardware. Muchos proyectos han intentado emular el tipo de aprendizaje jerárquico y reconocimiento de patrones que tiene lugar en la jerarquía neocortical, incluyendo mi propio trabajo en el campo de los modelos ocultos jerárquicos de Márkov. Partiendo de mi propia experiencia, una estimación conservadora es que emular un ciclo mediante un solo reconocedor de patrones del neocórtex biológico del cerebro requeriría de más o menos 3 000 cálculos. La mayoría de las simulaciones funcionan en una fracción de esta estimación. Como el cerebro funciona a unos $10^2$ (100) ciclos por segundo, se obtienen $3 \cdot 10^5$ (300 000) cálculos por segundo y por reconocedor de patrones. Dando por buena mi estimación de $3 \cdot 10^8$ (300 millones) de reconocedores de patrones, obtenemos unos $10^{14}$ (100 billones) de cálculos por segundo, un número congruente con mi estimación hecha en *La Singularidad está cerca**. En dicho libro predije que para simular funcionalmente el cerebro se necesitarían entre $10^{14}$ y $10^{16}$ cálculos por segundo (cps) y utilicé la cifra de $10^{16}$ cps para ser conservador. La estimación del experto en IA Hans Moravec, basada en la extrapolación de los requisitos computacionales del procesamiento visual inicial a lo largo del cerebro, es de $10^{14}$ cps, lo cual concuerda con lo dicho aquí.

Los ordenadores portátiles normales pueden alcanzar las $10^{10}$ cps, aunque usando recursos en la nube este nivel puede ser ampliamente amplificado. El superordenador más veloz, el japonés K Computer, ya ha alcanzado las $10^{16}$ cps (8). Como el algoritmo del neocórtex se encuentra masivamente repetido, la estrategia de utilizar chips neuromórficos tales como los del IBM SyNAPSE mencionados anteriormente también es prometedora.

En términos de requisitos de memoria, necesitamos unos 30 bits (más o menos cuatro bytes) por conexión para cubrir uno de los 300 millones de reconocedores de patrones. Si aceptamos la estimación de 8 inputs de media por cada reconocedor de patrones, obtenemos 32 bytes por reconocedor. Si añadimos un peso de un byte por input, obtenemos 40 bytes. Añádanse otros 32 bytes para las conexiones descendientes y se llega a los 72 bytes. Téngase en cuenta que la cifra de las ramificaciones hacia arriba y hacia abajo a menudo será mucho mayor que ocho, aunque estas grandísimas ramificaciones en árbol son compartidas por muchos reconocedores. Por ejemplo, puede que existan cientos de reconocedores

involucrados en el reconocimiento de la letra “p”. Estos engrosarán miles de dichos reconocedores en el siguiente nivel más alto, que se dedica a las palabras y frases que incluyen la “p”. Sin embargo, cada reconocedor “p” no repite el árbol de conexiones que engrosa todas las palabras y frases que incluyen la “p” (todas ellas comparten un árbol de conexiones de estas características). Lo mismo se cumple para las conexiones descendientes. Un reconocedor responsable de la palabra inglesa “APPLE” les dirá a todos los miles de reconocedores “E” a un nivel inferior al suyo que se está a la espera de una “E” si ya se ha visto aparecer “A”, “P”, “P” y “L”. Dicho árbol de conexiones no se repite en cada reconocedor de palabras o frases que quiera informar al siguiente nivel inmediatamente inferior de que se está a la espera de una “E”. Repito, estos árboles se comparten. Por esta razón, una estimación general de una media de ocho hacia arriba y ocho hacia abajo por cada reconocedor de patrones es una estimación razonable. Incluso si incrementamos esta estimación en concreto no cambia significativamente el orden de magnitud de la estimación resultante.

Con $3 \cdot 10^8$ (300 millones) reconocedores de patrones a 72 bytes cada uno, obtenemos un requisito de memoria general de unos $2 \cdot 10^{10}$ (20 mil millones) de bytes. De hecho, este es un número bastante modesto que ordenadores normales pueden superar hoy en día.

Estas estimaciones solo pretenden proporcionar estimaciones aproximadas del orden de magnitud requerido. Como los circuitos digitales son alrededor de 10 millones de veces más rápidos que los circuitos neocorticales biológicos, no necesitamos alcanzar el paralelismo del cerebro humano, un modesto procesamiento paralelo (comparado con los billones de pliegues en paralelo del cerebro humano) será suficiente. Así, podemos observar que los requisitos computacionales necesarios están siendo alcanzados. El recableado que realiza el propio cerebro sobre sí mismo (las dendritas que están creando continuamente nuevas sinapsis) también puede ser emulado mediante software usando enlaces, un sistema mucho más flexible que el método basado en la plasticidad que sigue el cerebro, el cual hemos visto que es impresionante pero limitado.

Ciertamente, la redundancia utilizada por el cerebro para conseguir sólidos resultados invariables puede ser replicada mediante emulaciones de software. Las matemáticas involucradas en la optimización de estos tipos jerárquicos de sistemas de aprendizaje autoorganizativos son bien conocidas. Además, la organización del cerebro está lejos de ser óptima. Por supuesto, no hay necesidad de que sea perfecta, basta con que sea lo suficientemente buena como para alcanzar el umbral que permite ser capaz de crear herramientas que pueden compensar las limitaciones propias.

Otra restricción a la que está sujeto el neocórtex humano es que no existe ningún proceso que elimine o por lo menos revise ideas contradictorias, lo cual explica por

qué a menudo el pensamiento humano es profundamente inconsistente. Poseemos un mecanismo débil para abordar esto, el llamado pensamiento crítico, pero sin embargo esta capacidad no se practica ni de lejos tanto como se debería. En un neocórtex basado en software, podríamos construir un proceso que detectara inconsistencias que tuvieran que ser revisadas de nuevo.

Es importante tener en cuenta que el diseño de una región completa del cerebro es más sencillo que el diseño de una sola neurona. Tal y como expuse anteriormente, los modelos suelen volverse más simples a un nivel superior (hágase una analogía con respecto a un ordenador). Tenemos que comprender detalladamente los procesos físicos de los simiconductores para modelizar un transistor, y las ecuaciones que subyacen tras un solo transistor son complejas. Un circuito digital que multiplique dos números requiere de cientos de ellos. Sin embargo, podemos modelizar este circuito de multiplicación muy fácilmente mediante una o dos fórmulas. Un ordenador entero que contiene miles de millones de transistores puede ser modelizado a través de su set de instrucciones y de la descripción de sus registros, que puede ser descrita en un puñado de páginas de texto y fórmulas. Los programas de software en un sistema operativo, los compiladores de lenguaje y los ensambladores son razonablemente complejos, pero la modelización de un programa en concreto, por ejemplo, un programa de reconocimiento del habla basado en la jerárquica modelización oculta de Márkov, puede también ser descrito en solo unas pocas páginas de ecuaciones. En ningún lugar de dicha descripción se encontrarían los detalles de los procesos físicos de los semiconductores o ni siquiera de la arquitectura del ordenador.

Una observación similar también es verdadera en el caso del cerebro. Un reconocedor neocortical de patrones en concreto que detecte una característica visual invariable en concreto (por ejemplo una cara) o que realice un filtrado del sonido a pasobanda que restrinja el input a un rango de frecuencias específico o que evalúe la proximidad temporal de dos acontecimientos, puede ser descrito mediante muchos menos detalles específicos que las relaciones físicas y químicas que controlan los neurotransmisores, los canales iónicos y otras variables sinápticas y dendríticas involucradas en los procesos neuronales. Aunque toda esta complejidad tiene que ser cuidadosamente considerada antes de avanzar al nivel conceptual inmediatamente superior, gran parte de ella puede ser simplificada a medida que los principios operacionales del cerebro sean revelados.

* "General problem solver" en el original.
* También conocido como el segundo teorema de Shannon.
* Título publicado por Lola Books.

CAPÍTULO NUEVE

# Experimentos mentales sobre la mente

Las mentes son simplemente lo que los cerebros hacen.

—MARVIN MINSKY, *THE SOCIETY OF MIND*

Cuando construyamos máquinas inteligentes no deberíamos sorprendernos si las encontramos tan confundidas y obstinadas en sus convicciones sobre la relación mente-materia, la consciencia, el libre albedrío, etc. como lo están los hombres.

—MARVIN MINSKY, *THE SOCIETY OF MIND*

## *¿Quién es consciente?*

La verdadera historia de la consciencia empieza con la primera vez que uno miente.

—JOSEPH BRODSKY

El único origen de la consciencia es el sufrimiento.

—FYODOR DOSTOEVSKY, *MEMORIAS DEL SUBSUELO*

Existe un tipo de planta que ingiere comida orgánica mediante sus flores. Cuando una mosca se posa sobre su flor, los pétalos se cierran sobre ella rápidamente y la sujetan hasta que la planta ha absorbido al insecto en su sistema. Sin embargo, no se cierran sobre nada que no sea una buena comida; ignorarán una gota de lluvia o un palo. Qué curioso que algo tan inconsciente tenga tan buen ojo para sus propios intereses. Si esto es ser inconsciente, ¿cuál es la utilidad de la consciencia?

—SAMUEL BUTLER, 1871

Hemos estado examinando el cerebro como entidad capaz de conseguir ciertos niveles de éxito. Sin embargo, dicha perspectiva deja a nuestros *yoes* fuera del cuadro general. Parece como si viviéramos en nuestros cerebros. Así, tenemos vidas subjetivas, pero ¿cuál es la relación entre la objetividad del cerebro que hemos expuesto hasta ahora, nuestros propios sentimientos y nuestra sensación de ser la persona que tiene experiencias?

El filósofo británico Colin McGinn (nacido en 1950) escribe que discutir "sobre la consciencia puede reducir incluso al pensador más meticuloso a un parloteo incoherente". La razón de esto es que a menudo la gente tiene opiniones no

contrastadas e inconsistentes sobre lo que el término consciencia significa exactamente.

Muchos analistas consideran que la consciencia es una forma de rendimiento, como por ejemplo la capacidad de autoreflexión, que es la capacidad para comprender los pensamientos propios y explicarlos. Esto yo lo describiría como la capacidad para pensar sobre el pensamiento de uno mismo. Presumiblemente, podríamos encontrar una manera de evaluar esta capacidad y luego utilizarla a modo de test para separar las cosas conscientes de las inconscientes.

Sin embargo, rápidamente tendríamos problemas al intentar implementar esta estrategia. Un bebé, ¿es consciente? ¿y un perro? La verdad es que no se les da muy bien describir su propio proceso de pensamiento. Hay personas que creen que los bebés y los perros no son seres conscientes precisamente porque no pueden explicarse a sí mismos. ¿Y qué hay del ordenador conocido como Watson? Se le puede poner en un modo en el que de hecho explica cómo encontró una contestación determinada. Dado que contiene un modelo de su propio pensamiento, ¿es por tanto Watson consciente mientras que el bebé y el perro no lo son?

Antes de ir más allá en el análisis de esta cuestión, es importante reflexionar sobre la distinción más significativa relacionada con ella: ¿Qué es lo que podemos averiguar a partir de lo que nos dice la ciencia, en contraposición con lo que sigue siendo una verdadera cuestión filosófica? Un punto de vista mantiene que la filosofía es una especie de posada a medio camino entre las cuestiones que todavía no se han sometido al método científico y las que sí lo han sido. Según esta perspectiva, una vez que la ciencia avanza lo suficiente como para resolver un conjunto de cuestiones en concreto, los filósofos pueden pasar a ocuparse de otros asuntos hasta que estos sean asimismo resueltos por la ciencia. Por tanto, para los defensores de este punto de vista, la cuestión de la consciencia se ha convertido en algo endémico, especialmente en lo que respecta a la cuestión: "¿Qué y quién es consciente?"

Consideremos estas declaraciones del filósofo John Searle: "sabemos que mediante ciertos mecanismos biológicos los cerebros dan lugar a la consciencia. […] Lo fundamental es tener claro que la consciencia es un proceso biológico como la digestión, la lactancia, la fotosíntesis o la mitosis. […] El cerebro es una máquina, más concretamente una máquina biológica, pero una máquina al fin y al cabo. De manera que el primer paso es comprender cómo actúa el cerebro para luego construir una máquina artificial que posea un mecanismo igualmente efectivo a la hora de producir consciencia" (1). Al leer esta cita, la gente suele sorprenderse, ya que dan por sentado que Searle está plenamente dedicado a la protección del misterio de la consciencia en oposición a reduccionistas como Ray Kurzweil.

El filósofo australiano David Chalmers (nacido en 1966) ha acuñado el término "el duro problema de la consciencia" para describir la dificultad de fijar este fundamental e indescriptible concepto. A veces, una breve frase encapsula toda una

escuela de pensamiento, de manera que se vuelve emblemática (por ejemplo, este es el caso de la frase de Hannah Arendt "la banalidad del mal"). Esto mismo le pasa a la famosa formulación de Chalmers.

Al describir la consciencia es muy fácil optar por considerarla según los atributos observables y medibles asociados con el hecho de tener consciencia; sin embargo, esta estrategia ignora la propia esencia de la idea. Acabo de mencionar el concepto de la metacognición, la idea de pensar sobre el propio pensamiento, como uno de dichos correlatos o versiones de la consciencia. Otros analistas mezclan inteligencia emocional o inteligencia moral con consciencia. Pero, de nuevo, nuestra capacidad de expresar un sentimiento de cariño, de entender un chiste o de ser atractivos son meras formas de rendimiento o de actuación, impresionantes e inteligentes si se quiere, pero facultades que no obstante pueden ser observadas y medidas (incluso en el caso de que no estemos de acuerdo en cómo definirlas). Comprender cómo el cerebro realiza este tipo de tareas y lo que sucede en el interior del cerebro cuando las realizamos es lo que configura la "sencilla" cuestión de la consciencia planteada por Chalmers. Por supuesto, este "sencillo" problema es todo menos eso, y representa quizá el desafío científico más difícil e importante de nuestra era. Por el momento, la "dura" cuestión de Chalmers es tan dura que entra esencialmente dentro de lo inefable.

Para apoyar esta distinción, Chalmers utiliza un experimento mental que tiene que ver con lo que él llama zombis. Un zombi es una entidad que actúa como una persona pero sin tener experiencias subjetivas. Es decir, un zombi no es consciente. Chalmers sostiene que dado que podemos concebir los zombis, al menos lógicamente estos tienen que ser posibles. Si estuviéramos en un cocktail y hubiera tantos humanos "normales" como zombis, ¿cómo los diferenciaríamos? Es posible que esto le recuerde a algún cocktail al que usted haya asistido.

Mucha gente responde a esta cuestión diciendo que ellos interrogarían a los individuos de los que quisieran asegurarse sobre sus reacciones emocionales ante sucesos e ideas. Según ellos, un zombi revelaría su falta de experiencias subjetivas mediante deficiencias en ciertos tipos de respuestas emocionales. Sin embargo, una respuesta de estas características simplemente falla a la hora de tomar en consideración los presupuestos del experimento mental. Si nos encontráramos con un persona no emocional (como por ejemplo un individuo con algún tipo de déficit emocional, como pasa en ciertos tipos de autismo), con un avatar o con un robot que no fuera tan convincente como lo es un ser humano emocional, entonces dicha entidad no sería un zombi. Recuerde que, según los presupuestos de Chalmers, un zombi *es* completamente normal en cuanto a su capacidad para responder, incluyendo la capacidad de reaccionar emocionalmente, lo único que pasa es que no tiene experiencias subjetivas. En definitiva no hay manera de identificar a un zombi

porque por definición no existe nada en su comportamiento que indique su naturaleza zombi. De manera que, ¿se trata de una distinción sin ninguna diferencia?

Chalmers no intenta responder esta difícil pregunta, pero sí que proporciona algunas posibilidades. Una es una forma de dualismo en la cual la consciencia per se no existe en el mundo físico, sino como una realidad ontológica separada. Según esta formulación, lo que una persona hace se basa en procesos de su cerebro. Como el cerebro es causalmente algo cerrado, podemos explicar las acciones de las personas por completo, incluyendo sus pensamientos, por medio de sus procesos. Por tanto, esencialmente la consciencia existe en otro reino, o por lo menos es un terreno separado del mundo físico. Esta explicación no permite que la mente (es decir, la propiedad consciente asociada al cerebro) afecte causalmente al cerebro.

Otra posibilidad que Chalmers toma en consideración no es lógicamente diferente de su noción del dualismo y a menudo recibe el nombre de panprotopsiquismo. Esta doctrina sostiene que todos los sistemas físicos son conscientes, aunque un humano es más consciente que, digamos, el interruptor de la luz. Ciertamente, yo estaría de acuerdo en que un cerebro humano posee más cosas de las que ser consciente que un interruptor de la luz.

Según mi opinión, que quizás pertenezca a una subescuela del panprotopsiquismo, la consciencia es una propiedad emergente de un sistema físico complejo. Desde esta perspectiva, un perro también es consciente, pero no tanto como un humano. Una hormiga posee también un cierto nivel de consciencia, pero mucho menor que el de un perro. Por otra parte, la colonia de hormigas podría considerarse como poseedora de un nivel de consciencia más alto que el de la hormiga individual, ya que ciertamente es más inteligente que una hormiga sola. Según esta estimación, un ordenador que emule satisfactoriamente la complejidad de un cerebro humano también poseería la misma consciente emergente que un humano.

Otra manera de conceptualizar la idea de la consciencia es como un sistema que posee “qualia”. ¿Qué son los qualia? Una definición del término es “experiencias conscientes”. Sin embargo, esto no nos lleva muy lejos. Considérese este experimento mental. Un neurocientífico es completamente daltónico, no el tipo de daltónico que confunde ciertos tonos de, por ejemplo, verde y rojo (como es mi caso), sino que sufre un daltonismo que le condena a vivir en un mundo completamente en blanco y negro. (En una versión más extrema de este escenario, el neurocientífico se ha criado en un mundo en blanco y negro y nunca ha visto ningún color, es decir, que en su mundo no hay colores). Sin embargo, el neurocientífico ha estudiado exhaustivamente los procesos físicos del color, sabe que la longitud de onda de la luz roja es de 700 nanómetros y conoce los procesos neurológicos de una persona que puede ver los colores de forma normal, por lo que sabe mucho sobre el modo en el que el cerebro procesa el color. Sabe más del color que la mayoría de la gente. Si quisiéramos ayudarla y explicarla cómo es de hecho la experiencia

asociada al “rojo”, ¿cómo lo haríamos?

Quizá le leeríamos una parte del poema “Rojo”, del poeta nigeriano Oluseyi Oluseun:

*Rojo el color de la sangre*
*símbolo de vida*
*Rojo el color del peligro*
*símbolo de muerte*

*Rojo el color de las rosas*
*símbolo de belleza*
*Rojo el color de los amantes*
*símbolo de unidad*

*Rojo el color del tomate*
*símbolo de buena salud*
*Rojo el color del caliente fuego*
*símbolo de deseo ardiente*

Esto le daría una idea bastante buena de algunas de las cosas que la gente asocia con el rojo, y quizá le permitiría realizar sus propias asociaciones durante una conversación sobre el color. (”Sí, amo el color rojo, tan caliente y fiero, tan peligrosamente hermoso […]”). Si quisiera, seguramente podría convencer a la gente de que ha experimentado el rojo, pero toda la poesía del mundo no la permitiría experimentarlo.

De forma similar, ¿cómo le explicaríamos a alguien que nunca ha entrado en contacto con el agua lo que se siente al bucear? De nuevo, nos veríamos obligados a recurrir a la poesía y sin embargo no hay forma de compartir la experiencia en sí. A estas experiencias es a las que nos referimos cuando hablamos de los qualia.

Muchos de los lectores de este libro habrán experimentado el color rojo. Sin embargo, ¿cómo puedo saber si su experiencia del rojo no es la misma experiencia que yo tengo cuando miro el azul? Ambos miramos un objeto rojo y con seguridad aseguramos que es rojo, y sin embargo esto no responde a la pregunta. Puede que experimente lo que usted experimenta cuando mira el azul, pero que ambos hayamos aprendido a decir que esas cosas son rojas. Nuevamente, podríamos proferir gran cantidad de poemas, pero solo reflejarían las cosas que la gente asocia con los colores y no expresarían la verdadera naturaleza de los qualia. De hecho, los daltónicos congénitos ya han leído mucho sobre los colores, ya que la literatura está repleta de referencias, y así es como experimentan de cierta manera los colores. ¿Cómo es su experiencia del rojo comparada con la experiencia de la gente con vista normal? En realidad se trata de la misma cuestión que la de la mujer* que vive en un mundo en blanco y negro. Es extraordinario que unos fenómenos tan habituales en nuestras vidas sean tan absolutamente inefables como para que sea imposible aseverar que estamos experimentando los mismos qualia.

Otra definición de qualia es el sentimiento de una experiencia. Sin embargo, esta definición no es menos circular que nuestros intentos anteriores por definir la consciencia, ya que las frases “sentimiento”, “tener una experiencia” y “consciencia” son meros sinónimos. La consciencia y la cuestión tan relacionada de los qualia representan una cuestión fundamental, o quizá la cuestión definitiva, de la filosofía (aunque la cuestión de la identidad puede que sea más importante, tal y como expondré en la última sección de este capítulo).

De manera que, en lo que respecta a la consciencia, ¿cuál *es* exactamente la cuestión una vez más? Es esta: ¿quién o qué es consciente? En el título de este libro hago referencia a la “mente” en vez de al “cerebro” porque una mente es un cerebro consciente. También podríamos decir que una mente posee libre albedrío e identidad. La propia aseveración de que estas cuestiones son filosóficas no es de por sí evidente. Así, yo mantengo que estas cuestiones nunca podrán ser completamente resueltas por la ciencia. En otras palabras, sin la aceptación de presuposiciones filosóficas, no existen experimentos falsables que podamos realizar para resolverlas. Si tuviéramos que construir un detector de consciencia, Searle querría estar seguro de que chorreara neurotransmisores biológicos. Por su parte, el filósofo norteamericano Daniel Dennett (nacido en 1942) se mostraría más flexible en cuanto al sustrato, pero es posible que quisiera saber si el sistema contiene o no un modelo de sí mismo y de su propio rendimiento. Esta visión se acerca más a la mía, pero en el fondo sigue tratándose de una asunción filosófica.

Regularmente se presentan propuestas que pretenden ser teorías científicas que enlazan la consciencia con algún atributo físico medible, a lo cual se refiere Searle como el “mecanismo causante de consciencia”. El científico, filósofo y anestesiólogo norteamericano Stuart Hameroff (nacido en 1947) ha escrito que “los filamentos citoesqueléticos son la raíz de la consciencia” (2). Se refiere a los finos hilos que se encuentran en todas las células (incluidas las neuronas, pero no solo en ellas) llamadas microtúbulos, las cuales dotan a las células de su integridad estructural y desempeñan un papel activo en la división celular. Sus libros y trabajos sobre esta cuestión contienen descripciones y ecuaciones detalladas que explican la posibilidad de que los microtúbulos jueguen un papel activo en el procesamiento de la información en el interior de la célula. Sin embargo, la conexión entre los microtúbulos y la consciencia requiere de un acto de fe que en lo fundamental no difiere del acto de fe implícito en una doctrina religiosa que describa un ser supremo que otorgue consciencia (a veces llamada “alma”) a ciertas entidades (normalmente humanas). La opinión de Hameroff se apoya en endebles evidencias, sobre todo en lo que se refiere a la observación de que los procesos neurológicos que podrían dar soporte a esta supuesta computación celular quedan detenidos durante la anestesia. No obstante, esto dista mucho de ser una evidencia convincente, ya que muchos

procesos se detienen durante la anestesia. Ni siquiera podemos asegurar que los sujetos no estén conscientes cuando están anestesiados. Todo lo que sabemos es que después las personas no recuerdan sus experiencias. Incluso esto no es algo universal, ya que algunas personas sí que recuerdan con precisión sus experiencias bajo anestesia, incluyendo por ejemplo conversaciones de los cirujanos. Bajo el nombre de percepción intraoperatoria, se calcula que este fenómeno ocurre unas 40 000 veces al año en los Estado Unidos (3). Pero incluso dejando esto a un lado, la consciencia y la memoria son conceptos completamente diferentes. Tal y como he expuesto en detalle, si me remonto a mis experiencias momento a momento durante el día de ayer, resulta que he experimentado un gran número de impresiones sensoriales, y sin embargo me acuerdo de muy pocas. ¿Significa esto que durante todo el día no fui consciente de lo que veía y oía? Se trata efectivamente de una buena pregunta cuya respuesta no es obvia.

El físico y matemático inglés Roger Penrose (nacido en 1931) hizo un acto de fe diferente al proponer que la fuente de la consciencia también se encontraba en los microtúbulos, aunque específicamente en sus supuestas capacidades de computación cuántica. Su razonamiento, aunque no lo diga explícitamente, parece ser el siguiente: la consciencia es algo misterioso y un fenómeno cuántico también lo es, de manera que de cierta forma tienen que estar relacionados.

Penrose comienza sus análisis con los teoremas de Turing sobre los problemas irresolubles y con el teorema de incompletud de Gödel relacionado con ellos. La premisa de Turing (que en el capítulo 8 fue expuesta con gran detalle) es que existen problemas algorítmicos que pueden ser planteados pero que no pueden ser resueltos mediante una máquina de Turing. Dada la universalidad computacional de la máquina de Turing, podemos llegar a la conclusión de que estos “problemas irresolubles” no pueden ser resueltos mediante ninguna máquina. El teorema de incompletud de Gödel proporciona un resultado similar en cuanto a la capacidad de demostrar conjeturas que involucren números. El argumento de Penrose es que el cerebro humano es capaz de resolver estos problemas irresolubles, de manera que es capaz de hacer cosas que una máquina determinista, como por ejemplo un ordenador, no puede. Por lo menos en parte, su motivación es la de elevar a los seres humanos por encima de las máquinas. Sin embargo, su premisa central (que los humanos pueden resolver los problemas irresolubles de Turing y Gödel) es, por desgracia, simple y llanamente falsa.

Un famoso problema irresoluble llamado el problema del castor atareado* dice así: encuéntrese el número máximo de unos que una máquina de Turing dotada de un cierto número de estados puede escribir en su cinta. Así, para determinar el castor atareado representado por el número $n$ construimos todas las máquina de Turing que tengan $n$ estados (que si $n$ es finito será un número finito) y luego determinamos el mayor número de unos que dichas máquinas escriban en sus cintas, exceptuando

aquellas máquinas de Turing que caigan en un bucle infinito. Esto es irresoluble porque a la vez que intentamos simular todos estos estados *n* de las máquinas de Turing, nuestro simulador cae en un bucle infinito al intentar simular una de las máquinas de Turing que caiga en un bucle infinito. Sin embargo, resulta que los ordenadores han sido capaces de determinar la función del castor atareado para ciertas *n*. Los humanos también, pero los ordenadores han resuelto el problema para muchas más *n* que los humanos no asistidos por ordenadores. Por lo tanto, podemos decir que por lo general los ordenadores son mejores que los humanos a la hora de resolver problemas irresolubles de Turing y Gödel.

Penrose enlazó estas supuestas capacidades transcendentes del cerebro humano con la computación cuántica que según su hipótesis tendría lugar en él. Según Penrose, estos efectos cuánticos neuronales serían en cierta manera intrínsecamente inalcanzables para los ordenadores, por lo que el pensamiento humano poseería una ventaja intrínseca frente a las máquinas. De hecho, los sistemas electrónicos normales utilizan efectos cuánticos, ya que los transistores se basan en el efecto túnel de los electrones para sobrepasar barreras, pero la computación cuántica del cerebro nunca ha sido demostrada. El rendimiento mental humano puede ser explicado satisfactoriamente mediante los métodos de computación clásicos, y en cualquier caso nada nos impide aplicar computación cuántica en los ordenadores. Ninguna de estas objeciones ha sido abordada por Penrose. El momento en el que Hameroff y Penrose unieron sus fuerzas fue cuando las voces críticas señalaron que el cerebro es un lugar demasiado cálido y desordenado como para realizar computación cuántica. Penrose dio con un vehículo perfecto en el interior de las neuronas que posiblemente podría servir de soporte para la computación cuántica, a saber, los microtúbulos sobre los que Hameroff había especulado como parte del procesamiento de la información en el interior de la neurona. Así, la tesis Hameroff-Penrose es que los microtúbulos de las neuronas realizan computación cuántica y que este hecho es responsable de la consciencia.

Esta tesis también ha sido criticada, por citar un ejemplo, por el físico y cosmólogo sueco-norteamericano Max Tegmark (nacido en 1967), quien llegó a la conclusión de que los eventos cuánticos en los microtúbulos solo podrían sobrevivir durante $10^{-13}$ segundos, lo cual es un lapso demasiado breve como para calcular resultados de importancia o como para afectar a los procesos neuronales. Existen ciertos tipos de problemas para los cuales la computación cuántica muestra capacidades superiores a las de la computación clásica, como por ejemplo la desencriptación de códigos a través de la factorización de números muy grandes. Sin embargo, el pensamiento humano no asistido por ordenadores ha demostrado ser muy malo en su resolución, y, en lo que se refiere a esta área, no puede competir ni siquiera con ordenadores clásicos, lo cual sugiere que el cerebro no muestra ninguna

capacidad de computación cuántica. Y no solo eso, incluso si un fenómeno como la computación cuántica existiese en el cerebro, no tendría que estar necesariamente relacionado con la consciencia.

## *Hay que tener fe*

> ¡Vaya una obra que es el hombre! ¡Qué noble es su raciocinio! ¡Qué infinitud de facultades! En su forma y movimiento, ¡qué expresivo y admirable! ¡Cómo se asemeja a un ángel en sus acciones! Su capacidad de comprensión, ¡como la de un dios! ¡La belleza del mundo! ¡El novamás de los animales! Y sin embargo yo me pregunto ¿qué es esta quintaesencia del polvo?
>
> —HAMLET, EN *HAMLET* DE SHAKESPEARE

Lo cierto es que todas estas teorías son actos de fe, y yo añadiría que en lo que concierne a la consciencia, el principio que ha de guiarnos es el que dice que "hay que tener fe"; es decir, que todos nosotros necesitamos hacer un acto de fe en lo que se refiere a la pregunta sobre qué y quién es consciente, así como sobre quién y qué somos en tanto en cuanto seres conscientes. De otra manera no podríamos levantarnos por la mañana. Sin embargo, tenemos que ser honestos sobre la necesidad básica que tenemos de hacer un acto de fe sobre este tema, y además tenemos que preguntarnos a nosotros mismos sobre el significado de nuestro acto de fe.

Las personas realizan actos de fe muy diversos, aunque pueda parecer lo contrario. Tras las presuposiciones filosóficas que cada uno hace sobre la naturaleza y origen de la consciencia subyacen desacuerdos sobre cuestiones que van desde los derechos de los animales hasta el aborto, y en el futuro se darán todavía más contiendas sobre lo que respecta a los derechos de las máquinas. Mi predicción sobre la materia es que en el futuro las máquinas darán la impresión de ser conscientes y que cuando hablen sobre sus qualia las personas biológicas las considerarán convincentes. Exhibirán actitudes que cubren todo el ancho rango de sutiles y conocidas emociones. Nos harán reír y llorar, y se enfadarán con nosotros si decimos que no creemos que sean conscientes. (También serán muy listas, de manera que no querremos que se enfaden). Acabaremos por aceptar que son personas conscientes. He aquí mi propio acto de fe: en cuanto las máquinas logren ser convincentes a la hora de hablar de sus qualia y experiencias conscientes, serán consideradas personas conscientes. He llegado a esta convicción a través de este experimento mental: imagínese que usted se encuentra con una entidad futura (un robot o un avatar) que es absolutamente convincente en lo que se refiere a sus reacciones emocionales. Se ríe convincentemente de sus chistes y además le hace reír y llorar (y no solamente pellizcándole). Le convence de su sinceridad al hablar de sus miedos y anhelos. Parece consciente en todas sus formas. De hecho, tiene el

aspecto de una persona. ¿Aceptaría usted a esta entidad como si fuera una persona consciente?

Si su reacción inicial es que le gustaría detectar algo que traicionara su naturaleza no biológica, entonces no está respetando los presupuestos de esta hipotética situación, ya que estos establecen que es completamente convincente. Una vez hecha esta presuposición, si esta entidad se enfrentara al peligro de su destrucción y reaccionara tal y como lo haría un humano mediante un sentimiento de terror, ¿reaccionaría usted igual de empáticamente que lo haría si presenciase una escena así en la que estuviera involucrado un humano? En mi caso, la respuesta es sí, y confío en que esta respuesta sería la misma en la mayoría o prácticamente todas las personas, independientemente de lo que a día de hoy declararan durante un debate filosófico. Repito: en lo que se hace énfasis es en la palabra "convincente".

Ciertamente, existen desacuerdos sobre la cuestión de si alguna vez nos encontraremos con una entidad así. La predicción que yo mantengo es que esto ocurrirá en 2029 y que se convertirá en algo habitual en la década de 2030. Pero dejando a un lado el marco temporal, creo que eventualmente acabaremos por considerar a estas entidades como conscientes. Considere cómo las tratamos ya cuando nos las encontramos bajo la forma de personajes de historias y películas; R2D2 de las películas de *Star Wars*, David y Teddy de la película *A.I.*, Data de la serie de televisión *Star Trek: The Next Generation*, Johnny 5 de la película *Corto Circuito*, Wall-E de la película de Disney *Wall-E*, T-800 (el Terminator bueno de la segunda y última películas de *Terminator*), Rachael La Replicante de la película *Blade Runner* (quien por cierto no es consciente de no ser humana), Bumblebee de la película, serie de televisión y comic *Transformers* y Sonny de la película *I, Robot*. Empatizamos con estos personajes pese a que sabemos que no son biológicos. Los consideramos personas conscientes, tal y como hacemos con los personajes biológicos humanos. Compartimos sus sentimientos y tememos por ellos cuando se encuentran en problemas. Si así es como a día de hoy tratamos a los personajes de ficción no biológicos es que así es como en la vida real trataremos a inteligencias futuras que no posean un sustrato biológico.

Si usted acepta el acto de fe que consiste en creer que una entidad no biológica convincente en sus reacciones relacionadas con los qualia es de hecho consciente, entonces considere lo que esto implica: que la consciencia es una propiedad emergente del patrón general de una entidad, no del sustrato en que se ejecuta.

Existe una laguna conceptual entre ciencia (que se corresponde con mediciones *objetivas* y con las conclusiones que podemos sacar de ellas) y consciencia (que es sinónimo de experiencia *subjetiva*). Obviamente, no podemos limitarnos a preguntar a la entidad en cuestión "¿eres consciente?" Si observáramos el interior de su "cabeza", sea esta biológica o no, para comprobarlo, tendríamos que hacer

presuposiciones filosóficas para determinar lo que estamos buscando. Por tanto, la pregunta sobre si una entidad es o no es consciente no es una cuestión científica. Basándose en esto, algunos analistas proceden a cuestionar si la propia consciencia tiene alguna base real. La escritora y filósofa inglesa Susan Blackmore (nacida en 1951) habla de la "gran ilusión de la consciencia". Reconoce la realidad del meme (idea) de la consciencia. En otras palabras, la consciencia existe como idea y existen muchísimas estructuras neocorticales que tienen que ver con esta idea, por no mencionar las palabras que han sido proferidas y escritas sobre ella. Sin embargo, no está claro que haga referencia a algo real. Blackburn continúa diciendo que ella no está necesariamente negando la realidad de la consciencia, sino que se limita a exponer los tipos de dilemas con los que nos encontramos al intentar fijar el concepto. Tal y como escribió el psicólogo y escritor británico Stuart Sutherland (1927–1998) en el *International Dictionary of Psychology*, "La consciencia es un fenómenos fascinante, pero evasivo. Es imposible especificar lo que es, lo que hace o el por qué de su evolución" (4).

Sin embargo, haríamos bien en no descartar este concepto demasiado alegremente como si se tratara de un mero debate entre filósofos (el cual, por cierto, se remonta a hace dos mil años en los diálogos de Platón). Tras la idea de la consciencia subyace nuestro sistema moral. A su vez, nuestro sistema legal está indirectamente construido sobre estas creencias morales. Si una persona acaba con la consciencia de alguien, como pasa por ejemplo en el asesinato, lo consideramos inmoral y, salvo algunas excepciones, un delito grave. Dichas excepciones también tienen que ver con la consciencia, ya que es posible que autoricemos que la policía o el ejército mate a ciertas personas conscientes para proteger a un mayor número de otras personas conscientes. Podemos debatir sobre la pertinencia de las excepciones concretas, pero el principio subyacente sigue siendo verdadero.

Atracar a alguien y hacer que sufra también se suele considerar inmoral e ilegal. Si destruyo mis posesiones, es probable que se trate de un hecho aceptable. Si destruyo sus posesiones sin su permiso, es probable que no sea un hecho aceptable, no porque esté haciendo sufrir a sus posesiones, sino a usted mismo como propietario. Por otro lado, si mis posesiones incluyen un ser consciente como es el caso de un animal, entonces yo como propietario del animal no tengo necesariamente la potestad moral o legal para hacer todo lo que se me antoje con él, ya que por ejemplo existen leyes contra la crueldad con los animales.

Como una gran parte de nuestros sistemas morales y legales están basados en la protección de la existencia y en la prevención del sufrimiento innecesario de las entidades conscientes, para hacer juicios responsables tenemos que contestar la pregunta sobre quién es consciente. Por tanto, la cuestión no es solo materia para el debate intelectual, como queda patente en la controversia que rodea a una cuestión como la del aborto. Debo señalar que la cuestión del aborto puede que sobrepase la

cuestión de la consciencia, ya que los activistas provida sostienen que el potencial de un embrión para en último término convertirse en una persona consciente es razón suficiente para que dicho embrión sea protegido, igual que alguien que está en coma también merece dicho derecho. Sin embargo, la cuestión fundamental de este debate versa sobre cuándo un feto se vuelve consciente.

Las percepciones sobra la consciencia también suelen afectar nuestros juicios sobre cuestiones controvertidas. Volviendo al tema del aborto, mucha gente distingue entre una medida como la píldora del día después, que previene la implantación del embrión en el útero durante los primeros días de embarazo, y el aborto en una etapa tardía. La diferencia tiene que ver con la probabilidad de que el feto de etapa tardía posea consciencia. Es difícil sostener que un embrión con una edad de tan solo unos días sea consciente, a no ser que se adopte una posición panprotopsíquica, pero incluso en estos términos estaría por debajo del animal más simple en términos de consciencia. De forma similar, reaccionamos de manera muy diferente cuando se trata del maltrato a los grandes simios y cuando se trata de, pongamos por caso, el maltrato a insectos. A día de hoy nadie se preocupa demasiado del dolor y sufrimiento infringidos a nuestro software informático, aunque sí que nos referimos profusamente a la capacidad del software para causar sufrimiento. Sin embargo, cuando el software del futuro tenga la inteligencia intelectual, emocional y moral de los humanos biológicos, este tema se convertirá en una verdadera preocupación.

Así, mi postura es la de aceptar que entidades no biológicas que sean completamente convincentes en sus reacciones emocionales serán personas, y mi predicción es que el consenso social también las aceptará. Nótese que esta definición va más allá de las entidades que puedan pasar el test de Turing, el cual exige dominar el lenguaje humano. Estas entidades son lo suficientemente parecidas a los humanos como para que yo las incluyera dentro del término persona, y creo que la mayor parte de la sociedad también lo haría, pero también englobo a entidades que evidencien reacciones emocionales como las humanas y no sean capaces de pasar el test de Turing, por ejemplo, niños pequeños.

¿Resuelve esto la cuestión filosófica sobre quién es consciente, por lo menos en lo que respecta a mí mismo y a las demás personas que hacen este acto de fe en particular? La respuesta es: *no del todo*. Solo hemos englobado un caso, el de las entidades que actúan de una manera humana. Aunque estamos discutiendo sobre entidades futuras no biológicas, estamos hablando de entidades que muestran reacciones convincentes similares a las humanas, de manera que esta postura es antropocéntrica. Sin embargo, ¿qué ocurre con otras formas de inteligencia extrañas que no se asemejen a los humanos? Podemos imaginarnos inteligencias que sean tan complejas, o incluso mucho más complejas e intrincadas, que las correspondientes a cerebros humanos, pero que sin embargo posean emociones y motivaciones

completamente diferentes a las nuestras. ¿Cómo decidir entonces si son o no conscientes?

Podríamos empezar por examinar a las criaturas del mundo biológico que tengan cerebros comparables a los de los humanos y sin embargo muestren tipos de comportamiento muy diferentes. El filósofo británico David Cockburn (nacido en 1949) ha escrito sobre el visionado de un vídeo en el que un calamar gigante era atacado (o por lo menos el calamar así lo pensaba. Cockburn lanzó la hipótesis de que este podría tener miedo de los humanos que portaba la videocámara). El calamar temblaba y se acongojaba, y Cockburn escribe: "respondía de una forma que inmediata y poderosamente me hizo pensar en que se trataba de miedo. Parte de lo impactante en esta secuencia era la manera en que se podía observar que el comportamiento de una criatura físicamente tan diferente a un ser humano se correspondía con algo tan poco ambiguo y tan específico como el miedo" (5). Cockburn llega a la conclusión de que el animal estaba sintiendo dicha emoción y sostiene que la mayoría de la gente que viera dicha película llegaría a la misma conclusión. Si aceptamos la descripción y conclusión de Cockburn como correctas, entonces tendríamos que añadir a los calamares gigantes a nuestra lista de entidades conscientes. Sin embargo, esto tampoco nos lleva muy lejos, ya que esto sigue basándose en nuestra reacción empática ante una emoción que reconocemos en nosotros mismos, lo cual sigue siendo una perspectiva ego o antropocéntrica.

Si nos salimos de la biología, la inteligencia no biológica será todavía más variopinta que la inteligencia del mundo biológico. Por ejemplo, algunas entidades puede que no teman su propia destrucción y puede que no necesiten tener las emociones que observamos en los humanos o en cualquier otra criatura biológica. Además, puede que haya inteligencias que pudiendo pasar el test de Turing no deseen intentarlo.

De hecho, hoy en día construimos robots que no poseen el sentido de la autoconservación para que así realicen misiones en medios peligrosos. No los consideramos todavía lo suficientemente inteligentes o complejos como para tener en cuenta sus sentimientos, pero es posible imaginar robots futuros de este mismo estilo que sean tan complejos como los humanos. ¿Entonces qué pasará con ellos?

Personalmente diría que si en el comportamiento de un dispositivo así yo observara un compromiso con un objetivo complejo y valioso, así como la capacidad de ejecutar decisiones y acciones importantes que le permitieran cumplir con su misión, me sentiría impresionado. Además, probablemente me molestaría que dicho dispositivo fuera destruido. Quizás esté estirando demasiado esta idea, ya que estoy dando cuenta de un comportamiento que no incluye muchas de las emociones a las que consideramos universales, ni en las personas, ni siquiera en criaturas biológicas de ningún tipo. Pero repito, estoy intentando poner en relación atributos que puedo encontrar en mí mismo y en otras personas. Después de todo, la idea de

una entidad totalmente entregada a una causa noble, la cual es llevada a cabo (o por lo menos es intentada) sin preocuparse del bienestar propio, no es algo completamente extraño en el ámbito de las experiencias humanas. En esta instancia también se encontraría una entidad que intentase proteger a los humanos biológicos o que de alguna manera intentase ayudar a nuestros fines.

¿Qué pasaría si esta entidad tuviera sus propios fines, estos fueran distintos de los de los humanos y no cumpliera con una misión que pudiéramos tildar de noble según nuestros propios términos? Entonces es posible que intentara saber si puedo relacionarme con dicha entidad, o por lo menos apreciar algunas de sus capacidades, de alguna otra forma. Si verdaderamente es muy inteligente, es probable que se le den bien las matemáticas, de manera que quizá pudiéramos conversar sobre eso. A lo mejor le gustan los chistes matemáticos.

Pero si la entidad no tiene ningún interés en comunicarse conmigo y yo no dispongo del acceso necesario a sus acciones y tomas de decisión como para que la belleza de sus procesos internos me conmueva, ¿significaría esto que la entidad no es consciente? Me veo obligado a llegar a la conclusión de que las entidades que no logren convencerme con sus reacciones emocionales, o que ni siquiera lo intenten, no son necesariamente carentes de consciencia. Sería difícil reconocer otra entidad consciente sin establecer un cierto nivel de comunicación empática, pero este juicio refleja mis propias limitaciones más que las de la entidad que está siendo considerada. Por lo tanto, tenemos que proceder con humildad. Ya es suficientemente complicado el ponernos a nosotros mismos en la piel subjetiva de otro ser humano como para no saber que esto será mucho más difícil en el caso de inteligencias que serán muy diferentes a la nuestra.

## *¿De qué es de lo que somos conscientes?*

> Si, a través del cráneo de una persona pensante y consciente, pudiéramos ver su cerebro, y si además el lugar de máxima excitabilidad se iluminara, entonces deberíamos ver, jugando sobre la superficie cerebral, un punto luminoso de bordes fantásticos y ondulantes que fluctuaría constantemente en tamaño y forma. Además estaría rodeado de una oscuridad más o menos densa que cubriría el resto del hemisferio.
>
> —IVAN PETROVICH PAVLOV, 1913 (6)

Volviendo a calamar gigante, podemos reconocer algunas de sus aparentes emociones, pero gran parte de su comportamiento es un misterio. ¿Qué se siente al ser un calamar gigante? ¿Qué se siente al estrujar un cuerpo invertebrado a través de una pequeña abertura? No poseemos el vocabulario necesario para responder a estas preguntas, ya que ni siquiera podemos describir las experiencias que compartimos con otras personas, como por ejemplo, la visualización del color rojo o la sensación

de nuestro cuerpo al ser salpicado por el agua.

Sin embargo, no tenemos que desplazarnos hasta el fondo del océano para encontrar misterios en la naturaleza de las experiencia conscientes, solo tenemos que tomar en consideración las nuestras. Por ejemplo, yo sé que soy consciente. Y asumo que usted, el lector, también lo es. (Sobre las personas que no han comprado mi libro no estoy seguro). ¿Pero *de qué* es de los lo que soy consciente? Es posible que usted se haga la misma pregunta.

Pruebe a hacer este experimento mental, que funcionará para todo aquel que sepa conducir un coche. Imagine que está conduciendo por el carril izquierdo de la autopista. Ahora cierre los ojos, agarre un volante imaginario y realice los movimientos para pasar al carril de su derecha.

Bien, antes de continuar leyendo, inténtelo.

He aquí lo que probablemente haya hecho: ha agarrado el volante, ha comprobado que el carril de la derecha esté despejado y, si así fuera, ha girado el volante hacia la derecha durante un corto periodo de tiempo. Después ha enderezado el volante. Y ya está.

Menos mal que no se encontraba en un coche de verdad, ya que habría atravesado todos los carriles de la autopista y se habría estrellado contra un árbol. Aunque seguramente debería haber mencionado que no debe intentar esto es un coche de verdad (aunque por el otro lado doy por asumido que usted tiene interiorizada la regla que dice que no se debe conducir con los ojos cerrados), no es este el problema fundamental en este caso. Si llevó a cabo el procedimiento que acabo de describir, y casi todo el mundo lo hace al llevar a cabo este experimento mental, lo ha hecho mal. Girar el volante hacia la derecha y luego enderezarlo hace que el coche se encamine en una dirección que es diagonal a la dirección original. Se pasará al carril de la derecha, tal y como se pretendía, pero se seguirá rodando hacia la derecha indefinidamente hasta salirse de la carretera. Lo que había que hacer al pasar con el coche al carril de la derecha era girar el volante hacia la izquierda tanto como se hubiera girado anteriormente hacia la derecha, y *después* enderezar. Esto hará que el coche vuelva a circular en línea recta hacia adelante sobre el nuevo carril.

Considere el hecho de que usted es un conductor habitual y de que usted ha realizado esta maniobra miles de veces. ¿Acaso no es consciente cuando lo hace? ¿Es que nunca ha prestado atención a lo que realmente estaba haciendo al cambiar de carril? Dando por hecho que usted no está leyendo este libro en el hospital mientras se recupera de un accidente al cambiar de carril, lo cierto es que usted domina esta capacidad. Y sin embargo no es consciente cuando la lleva a cabo, independientemente de las veces que haya realizado esta tarea.

Cuando la gente cuenta historias sobre sus experiencias, las describen como secuencias de situaciones y decisiones. Sin embargo, no es así como vivimos una

historia de primera mano. Nuestra experiencia original es una secuencia de patrones de alto nivel, algunos de los cuales pueden haber provocado ciertas sensaciones. Recordamos (si acaso) solamente un pequeño subconjunto de dichos patrones. Incluso si somos razonablemente precisos a la hora de rememorar una historia, utilizamos nuestra capacidad de fabulación para rellenar los detalles que faltan y hacer que la secuencia sea una historia coherente. No podemos estar seguros, partiendo de nuestra rememoración, de cómo fue nuestra experiencia consciente original, y sin embargo la memoria constituye el único acceso que tenemos a dicha experiencia. El momento presente es, por decirlo de alguna manera, fugaz, y rápidamente se convierte en un recuerdo o, más a menudo, no lo hace. Incluso si una experiencia se convierte en un recuerdo, esta se almacena, tal y como indica la PRTM, a modo de patrón de alto nivel compuesto de otros patrones pertenecientes a una jerarquía inmensa. Tal y como he señalado en varias ocasiones, casi todas las experiencias que tenemos (al igual que pasa con las ocasiones en las que cambiamos de carril) son olvidadas inmediatamente. De manera que verificar aquello que constituye nuestra propia experiencia consciente no es algo que esté a nuestro alcance.

## *El este está al este y el oeste está al oeste*

> Antes de los cerebros, en el mundo no había ni colores ni sonidos, ni tampoco sabores u olores, y probablemente ninguna sensación, así como ningún sentimiento o emoción.
>
> —ROGER W. SPERRY (7)

> René Descartes entra en un restaurante y se sienta a cenar. El camarero se acerca y le pregunta si le gustaría tomar un aperitivo.
>
> "No, gracias", dice Descartes, "me gustaría solamente pedir la cena".
>
> "¿Le gustaría que le dijera las especialidades del día?", pregunta el camarero.
>
> "No", dice Descartes impacientándose.
>
> "¿Le gustaría beber algo antes de cenar?", pregunta el camarero.
>
> Descartes, que es abstemio, se siente insultado. "¡Pienso que no!", dice indignado, y ¡POOF! desaparece.
>
> —UN CHISTE TAL Y COMO LO CUENTA DAVID CHALMERS*

Hay dos maneras de abordar las cuestiones que hemos estado considerando, la que sostiene occidente y la que sostiene oriente sobre la naturaleza de la consciencia y de la realidad. Desde la perspectiva occidental, se parte del mundo físico en el que evolucionan los patrones de información. Después de unos cuanto miles de millones de años de evolución, las entidades de dicho mundo evolucionan lo suficiente como para volverse seres conscientes. Según la perspectiva oriental, la consciencia constituye la realidad fundamental. Así, el mundo físico solo se hace realidad a

través de los pensamientos de los seres conscientes. En otras palabras, el mundo físico es la manifestación de los pensamientos de los seres conscientes. Por supuesto, esto no es más que una simplificación de filosofías complejas y diversas, pero representa las polaridades fundamentales de las filosofías de la consciencia y su relación con el mundo físico.

La división oriente-occidente sobre la cuestión de la consciencia también se manifiesta en escuelas de pensamiento opuestas en el campo de la física subatómica. Para la mecánica cuántica, las partículas existen a modo de los llamados campos de probabilidad. Cualquier medición realizada sobre ellos mediante cualquier dispositivo de medición provoca el llamado colapso de la función de onda, que significa que de pronto la partícula toma una localización determinada. Una opinión generalizada dice que dicha medición constituye una observación por parte de un observador consciente, ya que de lo contrario el concepto de medición no tendría sentido. Así, la partícula toma una localización determinada (así como otras propiedades tales como la velocidad) solo cuando está siendo observada. Básicamente, las partículas se imaginan que si nadie se toma la molestia de observarlas, no tienen que decidir el sitio en el que estar. Yo llamo a esto la escuela budista de la mecánica cuántica, porque según esta las partículas básicamente no existen hasta que no son observadas por una persona consciente.

Existe otra interpretación de la mecánica cuántica que evita esta terminología antropomórfica. Según este análisis, el campo que representa una partícula no es un campo de probabilidades, sino una mera función que posee valores diferentes en localizaciones diferentes. Por tanto, el campo representa fundamentalmente lo que la partícula es. Existen límites sobre los valores que se pueden dar en el campo dependiendo de las diferentes localizaciones, ya que el campo completo que representa una partícula representa solamente una cantidad de información limitada. De ahí proviene la palabra “cuanto”. El así llamado colapso de la función de onda, según esta interpretación, no es en absoluto un colapso. De hecho, la función de onda nunca desaparece. Lo único que ocurre es que el dispositivo de medición también está compuesto de partículas con campos y la interacción entre el campo de la partícula que está siendo medida y los campos de las partículas del dispositivo de medición da como resultado una medición en la que la partícula aparece en una localización determinada. Sin embargo, el campo sigue estando presente. Esta es la interpretación occidental de la mecánica cuántica, aunque es interesante reseñar que entre los físicos de todo el mundo la interpretación más extendida es la que he llamado interpretación oriental.

Hubo un filósofo cuya obra traspasó esta división entre oriente y occidente. El pensador anglo-austriaco Ludwig Wittgenstein (1889–1951) estudió la filosofía del lenguaje y del conocimiento y meditó sobre la cuestión de qué es lo que podemos llegar a saber realmente. Reflexionaba sobre este tema siendo un soldado en la

primera guerra mundial a la vez que tomaba notas para lo que sería el único libro que publicó en vida, *Tractatus Logico-Philosophicus*. La obra tenía una estructura inusual y solo gracias a los esfuerzos de su antiguo mentor, el matemático y filósofo británico Bertrand Russell, pudo encontrar editor en 1921. No obstante, se convirtió en la biblia de una de las escuelas filosóficas más importantes, el positivismo lógico. Esta escuela buscaba definir los límites de la ciencia. Así, el libro y el movimiento que surgió a su alrededor tuvieron influencia sobre Turing y sobre el nacimiento de la teoría de la computación y la lingüística.

El *Tractatus Logico-Philosophicus* anticipa la perspectiva de que todo conocimiento es intrínsecamente jerárquico. Así, el propio libro está organizado en proposiciones compartimentadas y numeradas. Por ejemplo, las primeras cuatro proposiciones del libro son:

1 El mundo es todo lo que acaece.
1.1 El mundo es la totalidad de los hechos, no de las cosas.
1.11 El mundo está determinado por los hechos y por ser todos los hechos.
1.12 Porque la totalidad de los hechos determina lo que acaece y también lo que no acaece.

Otra proposición importante del *Tractatus* de la que Turing se hizo eco es esta:

4.0031 Toda filosofía es una crítica del lenguaje.

En esencia, tanto el *Tractatus Logico-Philosophicus* como el movimiento del positivismo lógico aseveran que la realidad física existe de forma separada a la percepción que tenemos de ella. Sin embargo, todo lo que podemos conocer sobre ella es lo que percibimos mediante nuestros sentidos, que pueden ser aumentados mediante nuestras herramientas, y las inferencias lógicas que podemos realizar a partir de estas impresiones sensoriales. En el fondo, Wittgenstein trata de describir los métodos y fines de la ciencia. La última proposición del libro es la 7, “De lo que no se puede hablar, mejor es callarse”. En consecuencia, el primer Wittgenstein considera que la discusión sobre la consciencia es circular y tautológica, y por tanto una pérdida de tiempo.

Sin embargo, el Wittgenstein maduro rechaza esta posición por completo y dedica toda su atención filosófica a hablar sobre temas sobre los que anteriormente habría sostenido que había que guardar silencio. Sus escritos sobre esta revisión del pensamiento fueron reunidos y publicados en 1953, dos años después de su muerte, en un libro llamado *Investigaciones Filosóficas*. En él critica sus ideas de juventud expresadas en el *Tractatus* y las tilda de circulares y carentes de significado. Además, llega a la conclusión de que sobre lo que él había recomendado no hablar

era de hecho lo único sobre lo que merecía la pena reflexionar. Estos escritos tuvieron una gran influencia sobre los existencialistas y convierten a Wittgenstein en la única personalidad de la filosofía moderna que es artífice principal de dos escuelas de pensamiento a la vez de vanguardia y contradictorias en el campo de la filosofía.

¿Qué es lo que el pensamiento del Wittgenstein maduro considera como merecedor de reflexión y de comentario? Temas como la belleza y el amor, de los que dice que existen imperfectamente como ideas en las mentes de los hombres. Sin embargo, escribe que dichos conceptos existen en un ámbito perfecto e idealizado similar al de las "formas" perfectas que Platón describió en sus diálogos (otra obra que inauguró perspectivas aparentemente contradictorias en cuanto a la naturaleza de la realidad).

Un pensador que creo que ha sido malinterpretado es el filósofo y matemático francés René Descartes. Su famoso "pienso, luego existo" se suele interpretar como un elogio al pensamiento racional, en el sentido de que equivale a decir "pienso, es decir, que puedo llevar a cabo pensamiento lógico, por tanto merezco la pena". Así, a Descartes se le considera el arquitecto de la perspectiva racionalista occidental.

Sin embargo, al contemplar esta afirmación en el contexto de sus otros escritos, yo tengo una impresión diferente. Descartes estaba preocupado por lo que se conoce como el "problema cuerpo-mente". Concretamente, la pregunta es ¿cómo surge una mente consciente a partir de la materia física del cerebro? Desde esta perspectiva, parece que más bien estuviera intentando llevar al escepticismo racionalista hasta un punto de fractura, de manera que según mi opinión lo que la aseveración significa en realidad es que "pienso, que es lo mismo que decir que tiene lugar una experiencia subjetiva, luego de lo único de lo que podemos estar seguros es de que algo llamado *yo* existe". Descartes no podía estar seguro de que el mundo físico existiera porque de él solo disponemos de nuestras propias impresiones sensoriales, las cuales podrían ser falsas o completamente ilusorias. Sin embargo, sí que sabemos que la persona que sufre la experiencia existe.

Mi formación religiosa se produjo en una iglesia unitaria en la que estudiábamos todas las religiones del mundo. Nos pasábamos seis meses con, pongamos por caso, el budismo. Íbamos a misas budistas, leíamos sus libros y formábamos grupos de discusión junto con sus líderes. Después pasábamos a otra religión, como por ejemplo el judaísmo. El tema primordial era "muchos caminos hacia la verdad", además de la tolerancia y la transcendencia. Esta última idea significaba que el resolver las aparentes contradicciones entre las diferentes tradiciones no implicaba decidir que una era correcta y otra equivocada. Así, la verdad solo puede ser descubierta encontrando una explicación que supere (transcienda) las aparentes diferencias, especialmente en lo que respecta a cuestiones fundamentales de significado y propósito.

Así es como yo resuelvo la división entre oriente y occidente sobre la consciencia y el mundo físico. En mi opinión, ambas perspectivas han de ser ciertas.

Por una parte, es una insensatez negar el mundo físico. Incluso si viviéramos en una simulación, tal y como ha especulado el filósofo sueco Nick Bostrom, la realidad seguiría siendo un nivel conceptual que para nosotros sería real. Si aceptamos la existencia del mundo físico y la evolución que ha tenido lugar en él, entonces podemos observar que entidades conscientes han evolucionado a partir de él.

Por otra parte, la perspectiva oriental, que sostiene que la consciencia es lo fundamental ya que representa la única realidad verdaderamente importante, es asimismo difícil de negar. Consideremos solamente la alta estima en la que tenemos a las personas conscientes en contraposición a la que tenemos con las cosas inconscientes. A estas últimas no las consideramos como portadoras de ningún valor intrínseco excepto en la medida en que puedan influir sobre las experiencias subjetivas de las personas conscientes. Incluso si consideramos a la consciencia como una propiedad emergente dentro de un sistema complejo, no podemos sostener que se trata solamente de un atributo más, al mismo nivel que "la digestión" y "la lactancia" (para citar a John Searle). La consciencia representa lo verdaderamente importante.

La palabra "espiritual" suele usarse para hacer referencia a las cosas que poseen una significación fundamental o última. A mucha gente no le gusta utilizar la terminología procedente de las tradiciones espirituales o religiosas, ya que ello implica unos conjuntos de creencias con los que puede que no estén de acuerdo. Sin embargo, si dejamos a un lado las complejidades místicas de las tradiciones religiosas y nos limitamos a respetar "lo espiritual" como un término que implica algo cuyo significado para los humanos es muy profundo, entonces el concepto de consciencia sí que encaja, ya que refleja el valor espiritual último o fundamental. De hecho, la propia palabra "espíritu" suele ser usada para indicar consciencia.

Así, la evolución puede ser vista como un proceso espiritual en el que se crean seres espirituales, es decir, entidades conscientes. Asimismo, la evolución se desplaza hacia una mayor complejidad, un mayor conocimiento, una mayor inteligencia, una mayor belleza, una mayor creatividad y hacia la capacidad de expresar emociones de mayor transcendencia como el amor. Todas estas son descripciones que la gente ha utilizado para referirse al concepto de dios, aunque en estos aspectos a dios se le describe como carente de limitaciones.

La gente suele sentirse amenazada por discusiones que implican la posibilidad de que una máquina pueda ser consciente, ya que este tipo de consideraciones son interpretadas por ellos como denigrantes con respecto al valor espiritual de las personas conscientes. Sin embargo, esta reacción refleja una mala comprensión del

concepto de máquina. Tales críticas abordan la cuestión basándose en las máquinas que conocen a día de hoy, y aun con todo lo impresionantes que se están volviendo, estoy de acuerdo en que los ejemplos contemporáneos de la tecnología todavía no merecen que los respetemos como seres conscientes. Mi predicción es que se volverán indistinguibles de los humanos biológicos, a los cuales sí que consideramos seres conscientes, y que por tanto se beneficiarán del valor espiritual que otorgamos a la consciencia. No se trata de una denigración de las personas, sino de un aumento de nuestra comprensión sobre ciertas máquinas futuras. Es probable que para estas entidades tuviéramos que adoptar una terminología diferente, ya que serán un tipo diferente de máquinas.

De hecho, cuando hoy en día observamos el interior del cerebro y decodificamos sus mecanismos, descubrimos métodos y algoritmos que no solo podemos comprender, sino que también podemos recrear. Son "las partes de un molino que se impulsan las unas a las otras", para parafrasear al matemático y filósofo alemán Gottfried Wilhelm Leibniz (1646–1716) cuando se refería al cerebro en sus escritos. Los humanos ya representan una máquina espiritual. Y no solo eso, también convergeremos de una manera tan íntima con las herramientas que estamos creando que la distinción entre humano y máquina se irá difuminando hasta que la diferencia desaparezca. Dicho proceso ya está bastante avanzado, aunque la mayoría de las máquinas que suponen una extensión de nosotros mismos todavía no estén en el interior de nuestros cuerpos y cerebros.

# *Libre albedrío*

> Un aspecto fundamental de la consciencia es su habilidad para mirar hacia adelante, capacidad a la que llamamos 'prever'. Es la capacidad de planificar. En términos sociales, es la capacidad de describir un escenario que es probable que vaya a pasar, o que es posible que pase, en el contexto de interacciones sociales que todavía no han tenido lugar […]. Es un sistema a través del cual mejoramos nuestras posibilidades de hacer cosas que vayan en nuestro máximo beneficio. […] En mi opinión, el "libre albedrío" es nuestra aparente capacidad para elegir y actuar según aquello que nos parece útil y apropiado, así como nuestra insistencia en la idea de que dichas elecciones nos pertenecen a nosotros mismos.
>
> —RICHARD D. ALEXANDER

> ¿Debemos mantener que la planta no sabe lo que hace tan solo porque no posee ojos, ni oídos, ni cerebro? Si mantenemos que actúa única y exclusivamente de forma mecánica, ¿no nos veríamos obligados a admitir que muchas otras acciones aparentemente muy intencionadas también son mecánicas? A nosotros nos parece que la planta mata y se come la mosca mecánicamente, ¿es que no le puede parecer a la planta que el hombre mata y come ovejas mecánicamente?
>
> —SAMUEL BUTLER, 1871

> ¿Es el cerebro, cuya estructura es claramente doble, un órgano doble "aparentemente partido, pero que sin embargo representa la unión de una partición"?
>
> —HENRY MAUDSLEY (8)

La redundancia, tal y como hemos visto, constituye una estrategia fundamental del neocórtex. Sin embargo, en el cerebro existe otro nivel de redundancia, ya que sus hemisferios izquierdo y derecho, aunque no son idénticos, son en gran medida iguales. Al igual que ciertas regiones del neocórtex suelen acabar por procesar ciertos tipos de informaciones, los hemisferios, hasta cierto punto, también se especializan. Por ejemplo, el hemisferio izquierdo suele responsabilizarse del lenguaje verbal. Sin embargo, estas asignaciones también pueden enrutarse de forma diferente hasta el punto en que podemos sobrevivir y actuar de manera relativamente normal mediante solo una de las partes. Las investigaciones de las neuropsicólogas norteamericanas Stella de Bode y Susan Curtiss han dado cuenta de 49 niños que sufrieron hemisferoctomías (la extracción de una mitad de sus cerebros), una operación extrema que se realiza sobre pacientes que sufren enfermedades convulsivas que amenazan su vida, pero que solo se dan en un hemisferio del cerebro. A algunos de los pacientes que sufren este tratamiento les quedan secuelas y déficits, pero dichos déficits son muy específicos y los pacientes mantienen personalidades razonablemente normales. Muchos de ellos crecen sanos y el resto de las personas no se da cuenta de que les falta una mitad del cerebro. De Bode y Curtiss han escrito sobre niños a los que les falta el hemisferio izquierdo y que "desarrollan un lenguaje extraordinariamente bueno pese a la falta del 'hemisferio del lenguaje'" (9). Así, describen el caso de un estudiante que terminó sus estudios universitarios, hizo estudios de postgrado y que en los test de inteligencia estaba por

encima de la media. Los estudios muestran mínimos efectos secundarios a largo plazo en lo que respecta a la cognición en general, la memoria, la personalidad y el sentido del humor (10). En 2007, un estudio de los investigadores norteamericanos Shearwood McClelland y Robert Maxwell mostró en adultos resultados positivos similares también en el largo plazo (11).

Se ha hecho público que una niña alemana de 10 años nacida con solo una mitad del cerebro también es bastante normal. Incluso su visión en un ojo es casi perfecta, aunque los pacientes que sufren hemisferoctomías pierden parte del campo visual después de la operación (12). El investigador escocés Lars Muckli comentó: "La plasticidad del cerebro es asombrosa, pero nos sorprendimos mucho al ver lo bien que se ha adaptado el único hemisferio del cerebro de esta niña para compensar la mitad que falta".

Aunque ciertamente estas observaciones apoyan la idea de la plasticidad del neocórtex, lo más interesante es que implican que los humanos, aparentemente, tenemos dos cerebros, no uno, y que nos puede ir bastante bien en el caso de tener solo uno. Si perdemos un hemisferio, perdemos los patrones corticales que solo allí se encuentran almacenados, pero cada cerebro en sí es bastante completo. De manera que, ¿tiene cada hemisferio su propia consciencia? Existen razones para mantener que ese es precisamente el caso.

Consideremos los pacientes con cerebro dividido que poseen los dos hemisferios, pero que tienen cortado el canal que los une. El cuerpo calloso es un fardo de unos 250 millones de axones que conecta los hemisferios cerebrales izquierdo y derecho, y que les permite comunicarse y coordinarse entre ellos. Al igual que dos personas se pueden comunicar íntimamente entre ellas y actuar como una solo persona aunque estén físicamente separados y sean individuos diferentes, los dos hemisferios del cerebro también pueden funcionar como una unidad y seguir siendo independientes.

Como su propio nombre indica, en los pacientes con cerebro dividido el cuerpo calloso ha sido cortado o ha sufrido daños, cosa que deja a los pacientes con dos cerebros funcionales sin un nexo directo de comunicación entre ellos. En sus investigaciones, el psicólogo norteamericano Michael Gazzaniga (nacido en 1939) ha realizado exhaustivos experimentos sobre el pensamiento de cada hemisferio en pacientes con cerebro dividido.

El hemisferio izquierdo en dichos pacientes suele ver el campo visual derecho y viceversa. Gazzaniga y sus colegas mostraron al campo visual derecho de un paciente con cerebro dividido una foto de la pata de un pollo (que por tanto era vista por su hemisferio izquierdo) y una escena nevada al campo visual izquierdo (que por tanto era vista por su hemisferio derecho). Después se le mostró una serie de fotografías de manera que ambos hemisferios pudieran verlas. Se le pidió al paciente que escogiera una foto que tuviera relación con la primera fotografía. La mano

izquierda del paciente (controlada por su hemisferio derecho) señaló la fotografía de una pala, mientras que su mano derecha señaló la fotografía de un pollo. Hasta el momento todo iba bien, los dos hemisferios actuaban independientemente y con buen juicio. "¿Por qué has escogido esta?" preguntó Gazzaniga al paciente, el cual contestó verbalmente (bajo el control del centro lingüístico de su hemisferio izquierdo), "obviamente, la pata de pollo le corresponde al pollo". Pero entonces el paciente bajó la mirada y al notar que su mano izquierda señalaba la pala inmediatamente explicó (de nuevo bajo el control del centro lingüístico del hemisferio izquierdo) pero "se necesita una pala para limpiar el gallinero".

Esto es una confabulación. El hemisferio derecho, el que controla el brazo y la mano izquierdos, señalaba de forma correcta a la pala, pero como el hemisferio izquierdo, el que controla la respuesta verbal, no había percibido la nieve, se inventó una explicación, aunque no era consciente de estar fabulando. Se trata de asumir la responsabilidad de una acción que nunca decidió hacer y de hecho nunca hizo, pero que pensaba que sí había realizado.

Esto implica que cada uno de los dos hemisferios en un paciente de cerebro dividido posee su propia consciencia. Aparentemente, los hemisferios no se dan cuenta de que su cuerpo está controlado por dos cerebros, ya que aprenden a coordinarse entre ellos y sus decisiones son lo suficientemente ordenadas y coherentes como para que cada uno piense que las decisiones del otro son las suyas propias.

El experimento de Gazzaniga no demuestra que un individuo normal con un cuerpo calloso funcional tenga dos mitades cerebrales conscientes, pero sugiere dicha posibilidad. Aunque el cuerpo calloso permite una colaboración efectiva entre las dos mitades del cerebro, esto no significa necesariamente que no sean mentes separadas. Cualquiera de las dos podría ser engañada para que pensara que ha sido ella quien tomó todas las decisiones, ya que dichas decisiones serían lo suficientemente similares a las que hubiera tomado ella por sí misma. Después de todo, sí que ejerce mucha influencia en cada decisión mediante la colaboración con el otro hemisferio a través del cuerpo calloso. Así, a cada una de las dos mentes le parecería que es ella quien tiene el control.

¿Cómo comprobar la conjetura de que ambas partes son conscientes? Se podría evaluar cada una de las dos en busca de correlatos neurológicos de la consciencia, que es justo lo que Gazzaniga ha hecho. Sus experimentos demuestran que cada hemisferio actúa como un cerebro independiente. Así, la fabulación no se limita a los hemisferios cerebrales; todos y cada uno la llevamos a cabo de forma habitual. Cada hemisferio es más o menos tan inteligente como un humano, de manera que si creemos que un cerebro es consciente, entonces tenemos que concluir que cada hemisferio es consciente de forma independiente. Podemos evaluar los correlatos

neurológicos y podemos realizar nuestros propios experimentos mentales (por ejemplo, podemos considerar que si dos hemisferios cerebrales sin un cuerpo calloso funcional constituyen dos mentes conscientes por separado, entonces lo mismo tendría que ser cierto para dos hemisferios con un nexo funcional entre ellos), pero cualquier intento por detectar más directamente consciencia en cada hemisferio nos vuelve a enfrentar con la falta de una comprobación científica de la consciencia. Sin embargo, si aceptamos que cada hemisferio del cerebro es consciente, entonces ¿tenemos que admitir que la así llamada actividad inconsciente del neocórtex (que constituye el grueso de su actividad) también posee una consciencia independiente? ¿O quizá tenga más de una? De hecho, Marvin Minsky se refiere al cerebro como una "una sociedad de la mente" (13).

En otro experimento con cerebros divididos, los investigadores mostraban la palabra "campana" al cerebro derecho y "música" al cerebro izquierdo. Al paciente se le preguntaba qué palabra era la que veía. El centro del habla controlado por el hemisferio izquierdo decía "música". Entonces al sujeto se le mostraba una serie de fotografías y se le pedía que señalara la fotografía más relacionada con la palabra que acababa de mostrársele. Su brazo controlado por el hemisferio derecho señaló la campana. Al preguntarle por qué señalaba la campana, su centro lingüístico controlado por el hemisferio izquierdo respondió: "bueno, música, la última vez que escuché una música fue la de las campanas sonando fuera de este lugar". Dio esta explicación pese a que había otras fotografías que estaban mucho más relacionadas con la música.

De nuevo, se trata de una fabulación. El hemisferio izquierdo da explicaciones como si fuera suya una decisión que nunca ha tomado y nunca ha llevado a cabo. No lo hace por encubrir a un amigo (es decir, al otro hemisferio), es que realmente cree que la decisión fue suya.

Estas reacciones y decisiones pueden extenderse al campo de las respuestas emocionales. De forma que ambos hemisferios pudieran oírlo, los investigadores preguntaron a un paciente adolescente de cerebro dividido "¿Quién es tu mejor …?" y luego le comunicaron la palabra "amiga" solo a su hemisferio derecho a través del oído izquierdo. Gazzaniga recoge que el sujeto se ruborizó y se avergonzó, una reacción comprensible en un adolescente al que se le pregunta por su novia. Sin embargo, el centro del habla controlado por el hemisferio izquierdo indicó que no había oído ninguna palabra y pidió una clarificación: "¿Mi mejor qué?" Cuando se le volvió a pedir que respondiera a la pregunta, pero esta vez por escrito, la mano controlada por el hemisferio derecho escribió el nombre de su novia.

Las pruebas de Gazzaniga no son experimentos sobre el pensamiento, sino verdaderos experimentos mentales. Así, a la vez que ofrecen una interesante perspectiva sobre la cuestión de la consciencia, nos hablan más directamente sobre la cuestión del libre albedrío, ya que en todos estos casos un hemisferio cree que ha

tomado una decisión que en realidad nunca ha tomado. ¿Hasta qué punto es esto cierto en cuanto a las decisiones que tomamos en nuestro día a día?

Consideremos el caso de una paciente epiléptica de 10 años. El neurocirujano Itzhak Fried le realizó una operación cerebral mientras estaba despierta (lo cual es posible porque en el cerebro no hay receptores del dolor) (14). Cuando estimulaba un punto en concreto de su neocórtex, ella se reía. Al principio, el equipo médico pensó que podían estar provocando algún tipo de risa refleja, pero rápidamente se dieron cuenta de que lo que estaban activando era la propia percepción del humor. Aparentemente, habían encontrado un punto en el neocórtex (obviamente hay más de uno) que reconoce la percepción del humor. No solo se reía, sino que de hecho encontró que la situación era graciosa, aunque nada hubiera cambiado en cuanto a la situación más que el hecho de que habían estimulado este punto concreto de su neocórtex. Cuando la preguntaron por qué se reía, no contestó diciendo "oh, por nada en particular" o "es que acabáis de estimular mi cerebro", sino que inmediatamente se inventó una razón. Apuntaba a algo en la habitación y trataba de explicar por qué era gracioso. "Estáis tan graciosos ahí parados", solía decir.

Aparentemente, somos muy dados a explicar y racionalizar nuestras acciones, incluso cuando en realidad no fuimos nosotros quienes tomamos la decisión de llevarlas a cabo. De manera que, ¿cuál es nuestro grado de responsabilidad en nuestras decisiones? Consideremos los experimentos del profesor de psicología Benjamin Libet (1916–2007), de la *University of California at Davis*. Libet puso a los participantes frente a un cronómetro y les puso electrodos EEG en el cuero cabelludo. Les dio la orden de realizar tareas simples como presionar un botón o mover un dedo. A los participantes se les pidió que se fijaran en el tiempo reflejado en el cronómetro cuando "se dieran cuenta por primera vez del deseo o necesidad de actuar". Los tests indicaron un margen de error de solo 50 milisegundos en las evaluaciones hechas por los sujetos. También midieron una media de unos 200 milisegundos entre el momento en el que los sujetos anunciaron ser conscientes del deseo de actuar y el acto en sí (15).

Los investigadores también observaron las señales EEG procedentes de los cerebros de los sujetos. De hecho, la actividad cerebral involucrada en la iniciación de la acción por parte del córtex motor (que es responsable de llevar a cabo la acción) ocurre una media de aproximadamente 500 milisegundos antes de la realización de la tarea. Esto significa que el córtex motor está preparado para llevar a cabo la tarea alrededor de un tercio de segundo antes de que el sujeto ni siquiera sea consciente de que ha tomado la decisión de realizarla.

Las implicaciones de los experimentos de Libet han sido debatidas acaloradamente. El propio Libet llegó a la conclusión de que nuestra consciencia sobre la toma de decisiones parece ser una ilusión y que "la consciencia está al

margen". El filósofo Daniel Dennett a propósito comentó: "originariamente, la acción se precipita en alguna parte del cerebro, esto hace que las señales vuelen hasta los músculos haciendo una pausa en su ruta para decirle a usted, el agente consciente, lo que está pasando. Sin embargo, igual que todo buen oficial hace con un Presidente incompetente, se mantiene la ilusión de que usted fue quien propició todo" (16). Asimismo, Dennett ha cuestionado los tiempos registrados en el experimento. Básicamente viene a sostener que es posible que los sujetos no fueran conscientes de cuándo se volvían conscientes de la decisión de actuar. Podríamos preguntarnos: si el sujeto no es consciente de cuando se vuelve consciente de haber tomado una decisión, ¿entonces quién lo es? La cuestión está bien planteada. Tal y como he expuesto anteriormente, de aquello de lo que somos conscientes es algo que está lejos de ser evidente.

El neurocientífico indio-norteamericano Vilayanur Subramanian "Rama" Ramachandran (nacido en 1951) explica esta situación de una forma un tanto diferente. Dado que poseemos alrededor de 30 mil millones de neuronas en el neocórtex, en dicho lugar siempre están pasando muchas cosas y somos consciente de muy poco. Las decisiones, sean grandes o pequeñas, están siendo procesadas por el neocórtex constantemente y las posibles soluciones brotan en nuestra consciencia. En lugar de libre albedrío, Ramachandran sugiere que deberíamos hablar de "libre no-albedrío"*, es decir, del poder para rechazar soluciones propuestas por las partes no conscientes de nuestro neocórtex.

Tomemos como analogía una campaña militar. Los oficiales del ejército redactan una recomendación destinada al Presidente. Antes de recibir su aprobación, realizan los trabajos preliminares que permitirán que dicha decisión sea llevada a cabo. En un momento dado, la decisión propuesta llega hasta el Presidente, el cual da su aprobación, por lo que el resto de la misión se lleva a cabo. Como el "cerebro" representado en esta analogía engloba los procesos inconscientes del neocórtex (es decir, los oficiales subordinados al Presidente) así como sus procesos conscientes (el Presidente), además de observar actividad neuronal, también deberíamos ver acciones que se desarrollan antes de que la decisión del oficial sea tomada. Así, en cualquier situación siempre podemos discutir sobre el margen de libertad que los oficiales subordinados dieron al Presidente para que este aceptara o rechazara una recomendación (y ciertamente los Presidentes norteamericanos han hecho ambas cosas). Sin embargo, no debería sorprendernos que la actividad mental, incluso la del córtex motor, empezara antes de que nos diéramos cuenta de que una decisión había sido tomada.

Lo que enfatizan los experimentos de Libet es que existe mucha actividad en nuestros cerebros que subyace tras nuestras decisiones y de la que no somos conscientes. Sin embargo, ya sabíamos que de la mayor parte de lo que llega al neocórtex no somos conscientes. Por tanto, no debería sorprendernos que nuestras

acciones y decisiones surjan de actividades tanto inconscientes como conscientes. ¿Es esta una distinción importante? Si nuestras decisiones surgen de ambas actividades, ¿debería importarnos el separar las partes conscientes de las inconscientes? ¿no es cierto que ambos aspectos son representativos de nuestro cerebro? ¿no somos en último término responsables de todo aquello que ocurre en nuestro cerebro? "Sí, disparé a la víctima, pero no soy el responsable porque no estaba presentando atención" es seguramente una mala defensa. Aunque existen algunos angostos resquicios legales por los que una persona puede no ser considerada responsable de sus decisiones, por lo general tenemos que asumir la responsabilidad de todas las decisiones que tomamos.

Las observaciones y experimentos que he citado anteriormente constituyen experimentos mentales sobre la cuestión del libre albedrío, un tema que, al igual que la cuestión de la consciencia, ha sido debatido desde la época de Platón. El propio término "libre albedrío" data del siglo XIII, pero ¿qué significa realmente?

El diccionario Merriam-Webster lo define como la "libertad de los humanos para tomar decisiones que no vienen determinadas por causas anteriores o por intervención divina". Podrá observar que esta definición es absolutamente circular: "El libre albedrío es la libertad […]". Dejando a un lado la idea de la intervención divina como algo que se opone al libre albedrío, en esta definición existe un elemento útil, la idea de que una decisión "no [viene] determinada por causas anteriores". En un momento volveré sobre ello.

La *Stanford Encyclopedia of Philosophy* declara que el libre albedrío es la "capacidad de los agentes racionales para elegir el curso de una acción de entre varias alternativas". Sin embargo, según esta definición un mero ordenador posee libre albedrío, de manera que es todavía menos útil de la definición del diccionario.

De hecho, Wikipedia es un poco mejor. Define el libre albedrío como "la capacidad de los agentes para elegir libremente sin estar sujetos a ciertos tipos de restricciones […]. La restricción cuya influencia ha sido más importante ha sido […] el determinismo". Nuevamente se utiliza la palabra circular "libre" para definir el libre albedrío, pero por lo menos sí que se hace referencia a lo que se ha dado en llamar el principal enemigo del libre albedrío: *el determinismo*. A lo que la definición de Merriam-Webster citada más arriba también hace referencia es a las decisiones que "no están determinadas por causas previas".

De manera que, ¿qué es lo que queremos decir con la palabra determinismo? Si introduzco "2 + 2" en una calculadora y esta muestra "4", ¿puedo decir que la calculadora ha hecho uso de su libre albedrío al decidir mostrar un "4"? Nadie aceptaría esto como una demostración de libre albedrío, ya que la "decisión" estaba predeterminada por los mecanismos internos de la calculadora y por el input. Si introduzco un cálculo más complejo, seguimos llegando a la misma conclusión en lo

que respecta a su falta de libre albedrío.

¿Y qué pasa con Watson cuando responde a una pregunta de *Jeopardy!*? Aunque sus deliberaciones son mucho más complejas que las de una calculadora, muy pocos observadores (o ninguno) otorgaría libre albedrío a sus decisiones. Ningún ser humano sabe exactamente cómo funcionan estos programas, pero podemos identificar un grupo de gente que de forma conjunta puede describir la totalidad de sus métodos. Y lo que es más importante, su output viene determinado por (a) La totalidad de sus programas en el momento en el que la cuestión es planteada, (b) La propia cuestión, (c) El estado de sus parámetros internos que influyen sobre sus decisiones, y (d) El billón de bytes de sus bases de conocimiento que incluye enciclopedias enteras. Según estas cuatro categorías de información, su output está determinado. Podríamos especular sobre si plantear la misma cuestión conllevaría siempre la misma respuesta, pero Watson está programado para aprender de sus experiencias, de manera que existe la posibilidad de que contestaciones subsecuentes fueran diferentes. Sin embargo, esto no contradice el análisis, si no que más bien solo constituye un cambio en el punto (c), el parámetro que controla sus decisiones.

De manera que, ¿exactamente en qué se diferencia un humano de Watson para que al humano le asignemos libre albedrío pero no así al programa informático? Podemos identificar varios factores. Aunque Watson es mejor jugador de *Jeopardy!* que la mayoría (o la totalidad) de los humanos, sigue sin ser ni remotamente tan complejo como lo es un neocórtex humano. Watson posee muchos conocimientos y usa métodos jerárquicos, pero la complejidad de su pensamiento jerárquico sigue siendo considerablemente menor que la de un humano. Por lo tanto, ¿se reduce la diferencia a la escala de complejidad de su pensamiento jerárquico? Se puede esgrimir una razón para concluir que este no es el caso. En mi exposición sobre el tema de la consciencia hice referencia a que mi acto de fe particular es que estaría dispuesto a considerar a un ordenador que pasara un test de Turing válido como un ser consciente. Los mejores *chatbots* de hoy en día no son capaces de hacerlo (aunque están mejorando continuamente), por lo tanto mi conclusión con respecto a la consciencia tiene que ver con el nivel de rendimiento por parte de la entidad. Quizás esto también sea cierto en lo que se refiere a mi manera de otorgar libre albedrío.

Ciertamente la consciencia es una diferencia filosófica entre los cerebros humanos y los programas de software actuales. Consideramos a los cerebros humanos como conscientes, mientras que (*todavía*) no hacemos lo mismo con los programas de software. ¿Es este el factor subyacente tras el libre albedrío que estamos buscando?

Un sencillo experimento mental defendería que de hecho la consciencia es una parte vital del libre albedrío. Considérese una situación en la que alguien realiza una acción sin darse cuenta de que la está realizando, si no que es llevada a cabo por la

actividad inconsciente del cerebro de dicha persona. ¿Consideraríamos esto como una muestra de libre albedrío? La mayor parte de la gente respondería que no. Si la acción resultara ser dañina, seguramente seguiríamos responsabilizando a dicha persona de la acción e intentaríamos encontrar un acto consciente que pudiera haber hecho que dicha persona haya realizado acciones sin darse cuenta, como por ejemplo beber una copa de más o simplemente fracasar a la hora de entrenarse a uno mismo para tomar en consideración las decisiones antes de llevarlas a cabo.

Según algunos analistas, los experimentos de Libet son contrarios al libre albedrío ya que resaltan lo mucho que hay de inconsciente en nuestra toma de decisiones. Dado que entre los filósofos existe el consenso razonablemente aceptado de que el libre albedrío implica una toma de decisiones consciente, parece que esta es una condición previa al libre albedrío. Sin embargo, para muchos analistas la consciencia es una condición necesaria, pero no suficiente. Si nuestras decisiones (conscientes o no) están predeterminadas desde antes de que las realicemos, ¿cómo podemos sostener que nuestras decisiones son libres? Esta posición, que mantiene que el libre albedrío y el determinismo son incompatibles, se conoce bajo el nombre de incompatibilismo. Por ejemplo, el filósofo norteamericano Carl Ginet (nacido en 1932) sostiene que si los eventos del pasado, presente y futuro están determinados, entonces podemos decir que no tenemos ningún control sobre ellos o sus consecuencias. Nuestras supuestas decisiones y acciones son simplemente parte de esta secuencia predeterminada. Según Ginet, esto excluye el libre albedrío.

Sin embargo, no todo el mundo considera que el determinismo sea incompatible con el concepto de libre albedrío. Los compatibilistas sostienen que se es libre para decidir lo que se quiere aunque lo que se decida esté o pueda estar determinado. Por ejemplo, Daniel Dennett sostiene que aunque el futuro puede que venga determinado por el estado de cosas del presente, la realidad es que el mundo es tan intrincadamente complejo que no podemos saber lo que el futuro deparará. A lo que se refiere lo podemos llamar "expectativas". Así, ciertamente somos libres de realizar actos que difieren de dichas expectativas. Deberíamos sopesar cómo son nuestras decisiones y acciones comparadas con dichas expectativas, no con un futuro teoréticamente determinado que de hecho no podemos conocer. Según Dennett, esto es suficiente para defender la idea del libre albedrío.

Gazzaniga también adopta una posición compatibilista: "Personalmente, somos agentes responsables y tenemos que ser responsabilizados de nuestras acciones, aunque vivamos en un mundo determinista" (17). Un cínico podría interpretar está postura como sigue: usted no tiene control sobre sus acciones, pero de todas formas le vamos a echar la culpa de ellas.

Algunos pensadores descartan la idea del libre albedrío al considerarla como una ilusión. El filósofo escocés David Hume (1711–1776) lo describió como una

mera cuestión “verbal” caracterizada por “una sensación falsa o supuesta experiencia” (18). El filósofo alemán Arthur Schopenhauer (1788–1860) escribió que “a priori, todo el mundo se considera a sí mismo como completamente libre, incluso en lo que respecta a su acciones individuales, y cree que en cualquier momento puede comenzar a vivir de manera diferente. […] Sin embargo, a posteriori, gracias a la experiencia, se sorprende al comprobar que no es libre, sino que está sujeto a la necesidad y que pese a todas sus decisiones y reflexiones no es capaz de cambiar su conducta. Además descubre que desde el principio hasta el final de su vida va a tener que soportar su propia forma de ser, la cual él mismo desaprueba” (19).

Sobre esto yo añadiría varias cosas. El concepto de libre albedrío y de responsabilidad, una idea que está muy relacionada con él, es útil (de hecho es fundamental) para mantener el orden social independientemente de que exista o no el libre albedrío. Igual que claramente la consciencia existe a modo de meme, también lo hace el libre albedrío. Es posible que los intentos por probar su existencia, o incluso por definirlo, sean terriblemente circulares, pero lo cierto es que casi todo el mundo cree en dicha idea. Partes muy importantes de nuestro neocórtex de más alto nivel están dedicadas a la idea de que tomamos decisiones libremente y de que somos responsables de nuestras acciones. Independientemente de si en estricto sentido filosófico esto es cierto o posible, la verdad es que la sociedad sería mucho peor si no sostuviéramos dichas creencias.

Y no solo eso, además el mundo no está necesariamente determinado. Anteriormente he expuesto dos perspectivas diferentes sobre la mecánica cuántica que difieren con respecto a la relación que existe entre los campos cuánticos y el observador. Una interpretación muy extendida sobre la perspectiva basada en el observador asigna un papel determinado a la consciencia: las partículas no resuelven su ambigüedad cuántica hasta que no son observadas por un observador consciente. Así, existe otra controversia en la filosofía de los eventos cuánticos que tiene ver con nuestra discusión sobre el libre albedrío, una controversia que versa sobre la cuestión: ¿los eventos cuánticos están determinados o son aleatorios?

La interpretación más común sobre los eventos cuánticos es que cuando la función de onda en la que consiste una partícula “colapsa”, la ubicación de la partícula se vuelve única. Según esta interpretación, una gran cantidad de dichos eventos adoptará una distribución predecible (razón por la cual se considera que la función de onda es una distribución de probabilidades), pero la decisión que tome cada partícula que sufra un colapso de su función de onda es aleatoria. La interpretación opuesta es determinista, y sostiene que existe una variable oculta específica que no somos capaces de detectar por separado, pero cuyo valor determina la posición de la partícula. El valor o fase de la variable oculta en el momento en el que la función de onda colapsa determina la posición de la partícula.

La mayoría de los físicos cuánticos parece favorecer la idea de una decisión aleatoria según el campo probabilístico, pero las ecuaciones de la mecánica cuántica permiten la existencia de dicha variable oculta.

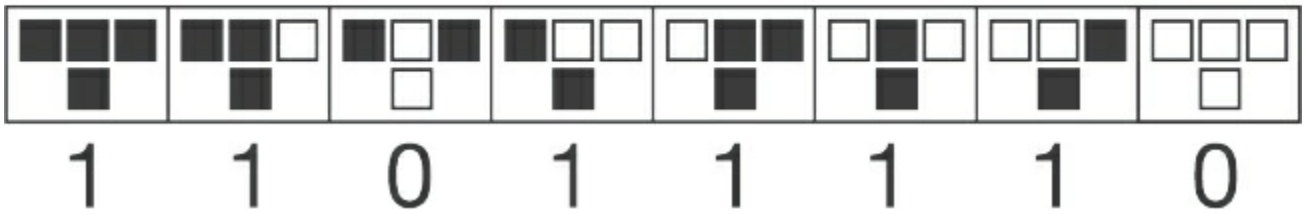
Aun así, es posible que el mundo, después de todo, no esté determinado. Según la interpretación de la probabilidad de la onda de la mecánica cuántica, existe una fuente constante de incertidumbre en el nivel más básico de la realidad. Sin embargo, esta observación no resuelve necesariamente las dudas de los incompatibilistas. Es cierto que según esta interpretación de la mecánica cuántica el mundo no está determinado, pero nuestro concepto de libre albedrío va más allá de las decisiones y de las acciones que son meramente aleatorias. La mayoría de los incompatibilistas opinarían que el concepto de libre albedrío también es incompatible con que nuestras decisiones sean esencialmente accidentales, ya que el libre albedrío parece implicar una toma de decisiones intencionada.

El Dr. Wolfram propone una manera de resolver este dilema. Su libro *Un nuevo tipo de ciencia* (2002) presenta de forma exhaustiva la idea de los autómatas celulares y su papel en cada faceta de nuestras vidas. Un autómata celular es un mecanismo en el cual el valor de celdas de información es recalculado continuamente como una función de las celdas que tiene cercanas. John von Neumann creó lo que en teoría es una máquina autoreplicante llamada constructor universal que quizá fuera el primer autómata celular.

El Dr. Wolfram ilustra su tesis mediante los autómatas celulares más sencillos posibles, un grupo de celdas en una línea unidimensional. En cada punto temporal cada célula puede tener uno de estos dos valores: negro o blanco. El valor de cada celda es recalculado en cada nuevo ciclo. El valor que toma una célula en el siguiente ciclo es una función de su valor actual, así como del valor de sus dos vecinos adyacentes. Cada autómata celular se caracteriza por una regla que determina la manera en la que se calcula si en el siguiente ciclo una celda será negra o blanca.

Consideremos el ejemplo de lo que el Dr. Wolfram llama la regla 222.

***Regla 222***

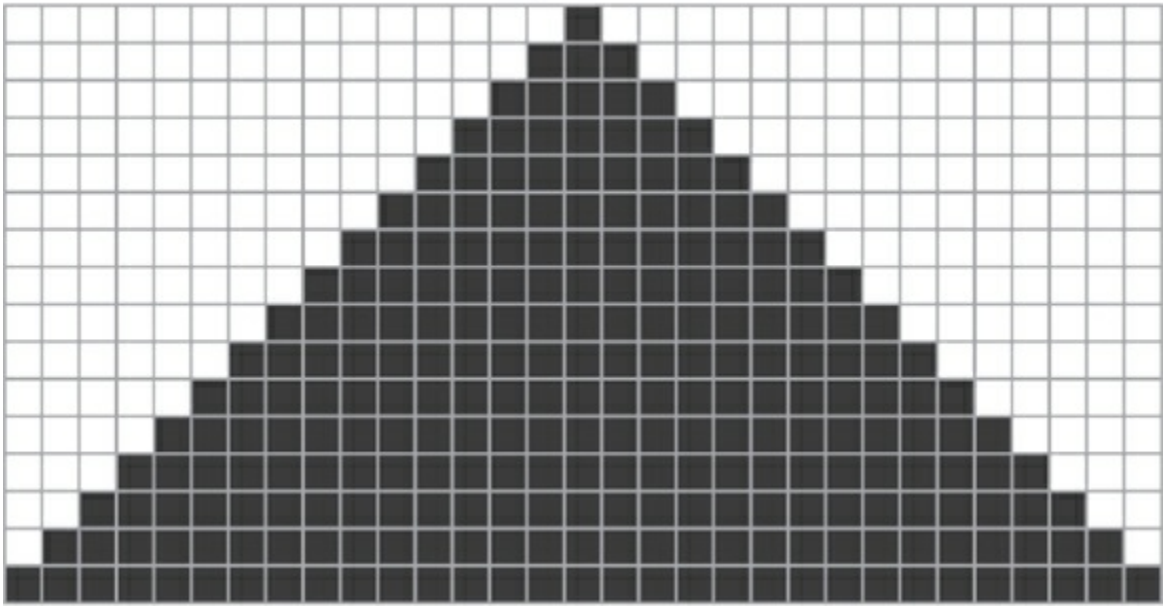



Las ocho posibles combinaciones para el valor de la celda que está siendo recalculada y para sus vecinos a izquierda y derecha se muestran en la fila de arriba. Su nuevo valor se muestra en la fila de abajo. De manera que, por ejemplo, si la

celda es negra y sus dos vecinos también son negros, entonces la célula seguirá siendo negra en la siguiente generación (véase la subregla más a la izquierda de la regla 222). Si la celda es blanca, su vecino izquierdo es blanco y su vecino derecho es negro, entonces la celda cambiará a negro en la siguiente generación (véase la subregla de la regla 222 que está colocada la segunda según se empieza por la derecha).

El universo de este sencillo autómata celular consta tan solo de una fila de celdas. Si comenzamos con tan solo una celda negra en el medio y observamos la evolución de las celdas durante varias generaciones, en las que cada fila según descendemos representa una nueva generación de valores, el resultado de la regla 222 tiene este aspecto:

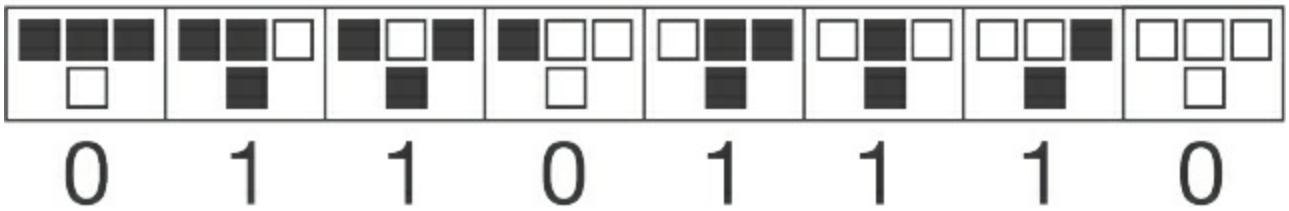
***Regla 222***



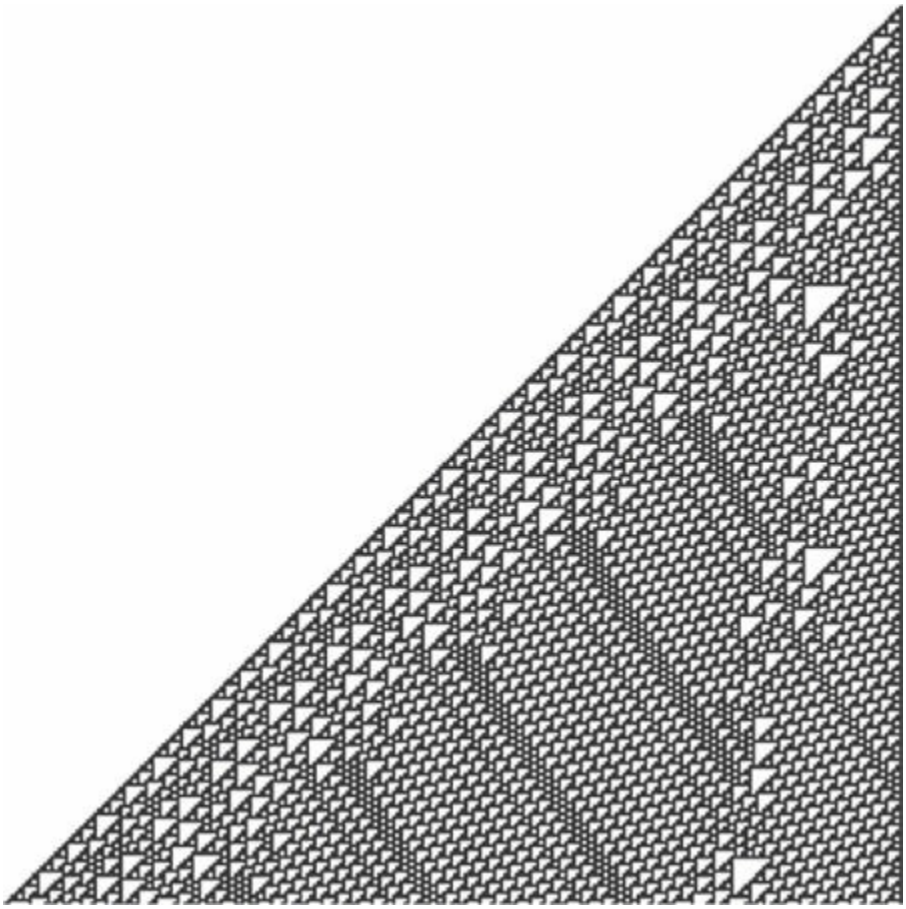
Un autómata se basa en una regla y una regla define si la celda será negra o blanca según los ocho posibles patrones que existen en la generación actual. Así, existen $2^8 = 256$ posibles reglas. El Dr. Wolfram ha ordenado todas las 256 reglas posibles de un autómata y a cada una le ha asignado un código que va desde el 0 hasta el 255. Curiosamente, estas 256 máquinas teóricas poseen propiedades muy diferentes. Los autómatas de la clase que el Dr. Wolfram llama I, como por ejemplo la regla 222, crean patrones muy predecibles. Si nos preguntáramos cuál era el valor de la celda del medio después de un billón de billones de iteraciones de la regla 222, podríamos contestar fácilmente: negro.

Sin embargo, los autómatas de clase IV son mucho más interesantes y vienen ilustrados por la regla 110.

## *Regla 110*




Muchas generaciones de este autómata tienen este aspecto:



Lo interesante del autómata de la regla 110 y de los autómatas clase IV en general es que los resultados son completamente impredecibles. Los resultados pasan los tests matemáticos de aleatoriedad más estrictos y sin embargo no se limitan a generar ruido. Hay patrones que se repiten, pero lo hacen de manera extraña e impredecible. Si nos preguntáramos por el valor de una celda en concreto después de un billón de billones de iteraciones, no habría manera de responder dicha cuestión sin que la máquina recorriera todas y cada una de las diferentes generaciones. Obviamente, la solución está determinada, ya que se trata de una máquina determinista muy simple, pero es completamente impredecible si no se hace funcionar la propia máquina.

La tesis principal del Dr. Wolfram es que el mundo es un gran autómata celular de clase IV. La razón por la cual su libro se titula *Un nuevo tipo de ciencia* es que

esta teoría contradice la mayoría del resto de leyes científicas. Si un satélite orbita La Tierra, podemos predecir dónde estará en cinco años sin tener que pasar por cada momento de un proceso simulado en el que usáramos la ley de la gravedad y calculáramos el lugar en el que estará durante los diferentes momentos del futuro lejano. Sin embargo, los estados futuros de los autómatas celulares de clase IV no pueden ser predichos sin simular cada paso del camino. Si el universo es un autómata celular gigante, tal y como postula el Dr. Wolfram, no habría un ordenador lo suficientemente grande que pudiera ejecutar una simulación así, ya que todo ordenador sería un subconjunto del universo. Por lo tanto, el estado futuro del universo es completamente incognoscible aunque esté determinado.

Así, aunque nuestras decisiones estén determinadas porque nuestros cuerpos y cerebros son parte de un universo determinista, sin embargo son intrínsecamente impredecibles, ya que vivimos en (y formamos parte de) un autómata clase IV. No podemos predecir el futuro de un autómata clase IV si no dejamos que el futuro tenga lugar. Según el Dr. Wolfram, esto es suficiente como para que se dé el libre albedrío.

No tenemos que recurrir al universo para encontrar acontecimientos futuros que están determinados pero son impredecibles. Ningún científico que haya trabajado en Watson puede predecir lo que hará porque el programa es demasiado complejo y diverso, y su rendimiento se basa es un conocimiento que es demasiado amplio como para que cualquier humano pueda abarcarlo. Si nos creemos que los humanos hacen uso de su libre albedrío, entonces esto implica que tenemos que permitir que versiones venideras de Watson o de máquinas similares a Watson también lo ejerciten.

Mi propio acto de fe es que creo que los humanos poseen libre albedrío, y como actúo como si ese fuera el caso, me encuentro fuertemente condicionado a la hora de encontrar ejemplos entre mis propias decisiones que ilustren este hecho. Consideremos la decisión de escribir este libro. Yo nunca tomé dicha decisión. Por el contrario, la idea del libro lo decidió por mí. Por lo general, me encuentro a mi mismo a merced de ideas que parecen implantarse por sí solas en mi neocórtex y toman el control sobre él. ¿Y qué pasa con la decisión de casarme, decisión que tomé (en colaboración con otra persona) hace 36 años? Hasta entonces seguí el programa habitual. Primero me sentí atraído por una chica guapa. Eso me llevó a pretenderla, y luego me enamoré. ¿Dónde queda el libre albedrío en todo esto?

¿Y qué hay de las pequeñas decisiones que tomo diariamente, por ejemplo las palabras concretas que elijo para escribir mi libro? Empiezo con un pedazo de papel en blanco virtual. Nadie me dice lo que tengo que hacer. Ningún editor mira sobre mi hombro. Mis elecciones dependen *por completo* de mi mismo. Soy libre (*totalmente libre*) de escribir *cualquier cosa* que yo…

*grok…*

*¿Grok?* Vale, lo hice, finalmente utilicé mi libre albedrío. Iba a escribir la palabra "quiera", pero tomé la decisión libre de escribir en su lugar algo totalmente inesperado. Quizá sea la primera vez que he tenido éxito a la hora de ejercer mi libre albedrío en estado puro.

O no.

Debería resultar evidente que esto no ha sido una muestra de mi voluntad, sino de intentar explicar algo (y quizás de exhibir un pobre sentido del humor).

Aunque comparto con Descartes su confianza en que soy consciente, no estoy tan seguro sobre el libre albedrío. Es complicado escapar a la conclusión de Schopenhauer de que "podemos hacer lo que deseamos, pero en todo momento de nuestras vidas podemos *desear* solamente una cosa en concreto y absolutamente nada más que dicha cosa" (20). No obstante continuaré actuando como si tuviera libre albedrío y continuaré creyéndomelo mientras no tenga que explicar por qué.

## *Identidad*

> Una vez un filósofo tuvo el siguiente sueño.
>
> Primero aparecía Aristóteles y el filósofo le decía "¿Podrías hacerme un resumen de quince minutos sobre toda tu filosofía?"
>
> Ante la sorpresa del filósofo, Aristóteles le hizo una excelente exposición en la que comprimió una gran cantidad de material en tan solo 15 minutos. Pero entonces el filósofo esgrimió una determinada objeción ante la cual Aristóteles no pudo contestar. Confundido, Aristóteles desapareció.
>
> Entonces apareció Platón y se volvió a repetir la misma situación. La objeción que le hacía el filósofo a Platón era la misma que le había hecho a Aristóteles. Platón tampoco pudo responderla y desapareció.
>
> A continuación aparecieron uno a uno todos los filósofos famosos de la historia y el filósofo les refutaba a todos y cada de ellos uno mediante la misma objeción.
>
> Después de que el último filósofo se desvaneciera el filósofo se dijo a sí mismo "Sé que estoy dormido y que todo esto lo estoy soñando. Sin embargo, ¡he encontrado una refutación universal para todos los sistemas filosóficos! Probablemente, mañana cuando me despierte lo habré olvidado ¡y entonces el mundo sí que se habrá perdido algo importante!" Con voluntad de hierro, el filósofo se obligó a sí mismo a despertar, se precipitó sobre su escritorio y escribió su refutación universal.
>
> Entonces se volvió a meter en la cama con un suspiro de alivio. La mañana siguiente, al despertar, fue hasta su escritorio y vio lo que había escrito. Ponía: "eso lo dirás *tú*".
>
> —RAYMOND SMULLYAN, CITADO POR DAVID CHALMERS (21)

Sobre lo que todavía me pregunto más que si sobre soy o no consciente, o sobre si tengo libre albedrío, es por qué soy consciente de las experiencia y decisiones de esta persona en concreto que escribe libros, se divierte haciendo senderismo y montando en bicicleta, toma suplementos nutricionales, etc. Una respuesta obvia sería "porque ese eres tú".

Probablemente dicho intercambio no sea más tautológico que mis contestaciones anteriores ante preguntas sobre la consciencia y el libre albedrío. Sin embargo, tengo una respuesta mejor para responder por qué mi consciencia está asociada a esta

persona en concreto: esto se debe a que yo me he creado a mí mismo para ser quien soy.

Un conocido aforismo dice que “somos lo que comemos”. Todavía más cierto es decir “somos lo que pensamos”. Tal y como he expuesto, todas las estructuras jerárquicas de mi neocórtex que definen mi personalidad, capacidades y conocimientos son el resultado de mis propios pensamientos y experiencias. Las personas con las que decido interactuar, así como las ideas y proyectos en los que elijo involucrarme, son todos factores decisivos de en quién me he convertido. Por lo tanto, lo que como también refleja las decisiones tomadas por mi neocórtex. Aceptando por el momento el lado positivo de la dualidad sobre el libre albedrío, son mis propias decisiones las que dan como resultado la persona que soy.

Independientemente de cómo hemos acabado siendo quienes somos, todos deseamos que nuestra identidad se perpetúe. Si usted no poseyera el deseo de sobrevivir, no estaría leyendo este libro. Toda criatura persigue este objetivo, que es el factor decisivo de la evolución. La cuestión de la identidad es quizá más complicada de definir que la de la consciencia o el libre albedrío, pero podría decirse que además es más importante. Después de todo, necesitamos saber quiénes somos si pretendemos preservar nuestra existencia.

Consideremos este experimento mental: usted se encuentra en el futuro con tecnologías más avanzadas que las de hoy. Mientras duerme, un grupo de personas escanea su cerebro y recoge todo detalle importante en él. Quizá lo hagan mediante una máquina de escaneado del tamaño de células sanguíneas que se desplazan por los capilares de su cerebro o mediante alguna otra sutil tecnología no invasiva. Así, estás personas poseen toda la información referente a su cerebro en un determinado momento. También recogen y registran cualquier detalle corporal que pueda reflejar su estado mental, como por ejemplo el sistema endocrino. Entonces ubican este “fichero mental” en un cuerpo no biológico que parece y se mueve como usted y posee la necesaria sutileza y agilidad para hacerse pasar por usted. Por la mañana le informan sobre esta transferencia y se encuentra (quizás sin ser visto) con su clon mental, al que llamaremos Usted 2. Usted 2 está hablando sobre su vida como si fuera usted y contando cómo esa misma mañana ha descubierto que le ha sido dada una versión 2.0 de su cuerpo mucho más resistente que la original. “¡Oye, como que me gusta este nuevo cuerpo!” exclama.

La primera cuestión a considerar es: ¿Es Usted 2 consciente? Ciertamente parece serlo. Ha pasado el test que he explicado anteriormente, ya que posee las sutiles formas que le permiten ser una persona que siente y es consciente. Si usted es consciente, entonces Usted 2 también lo es.

Si usted desapareciera nadie lo notaría. Usted 2 iría por ahí diciendo ser usted. Todos sus amigos y seres queridos estarían satisfechos con la situación y quizás alegres porque ahora posee un cuerpo y un sustrato mental más resistentes que los

que solía tener. Quizás sus amigos con inquietudes más filosóficas expresarían su preocupación, pero por lo general todo el mundo se alegraría, incluido usted, o por lo menos la persona que tan convincentemente dice ser usted.

De manera que su antiguo cuerpo y cerebro ya no son necesarios, ¿no es cierto? ¿Entonces podemos desecharlos?

Probablemente usted no esté de acuerdo con esto. He indicado que el escáner no era invasivo, por lo que usted sigue estando presente y con consciencia. Además, su sentido de la identidad todavía le acompaña, no como en el caso de Usted 2, aunque Usted 2 piense que es una continuación de usted. Puede que Usted 2 ni siquiera sepa que usted existe o que existió. De hecho, usted tampoco conocería de la existencia de Usted 2 si no se lo hubieran comunicado.

¿A qué conclusión llegamos? Usted 2 es consciente, pero no es la misma persona que es usted, Usted 2 tiene una identidad diferente. Es muy similar a usted, mucho más de lo que lo sería un clon genético, ya que también comparte todos sus patrones neocorticales y todas sus conexiones. Aunque la verdad es que debería decir que compartió dichos patrones en el momento en que fue creado. En dicho instante, ustedes dos empezaron a ir por su cuenta (neocorticalmente hablando). Usted sigue estando presente. No está experimentando lo mismo que Usted 2. Conclusión: Usted 2 no es usted.

De acuerdo hasta el momento. Consideremos ahora otro experimento mental, uno que considero más realista en términos de lo que deparará el futuro. Usted es operado y se le remplaza una parte muy pequeña de su cerebro por una unidad no biológica. Usted está convencido de que esto es seguro y existen noticias de los muchos beneficios que reporta.

Hoy esto ya no es tan inverosímil, ya que se realiza de forma rutinaria en personas con impedimentos neurológicos y sensoriales, tal y como es el caso de los implantes neuronales para la enfermedad de Párkinson y de los implantes de cóclea para sordos. En estos casos, el dispositivo computerizado es colocado en el interior del cuerpo pero fuera del cerebro, aunque conectado a él (o en el caso de los implantes de cóclea al nervio auditivo). En mi opinión, el hecho de que el ordenador esté colocado físicamente fuera del cerebro no es filosóficamente relevante, ya que de hecho estamos mejorando el cerebro y remplazando mediante un dispositivo computerizado aquellas funciones que no puede realizar adecuadamente. En la década de 2030, cuando inteligentes dispositivos computerizados tengan el tamaño de células sanguíneas (y tenga en cuenta que los glóbulos blancos son células lo suficientemente inteligentes como para reconocer y combatir patógenos), los introduciremos de forma no invasiva, sin necesidad de cirugía.

Volviendo a nuestro escenario futuro, usted ha sufrido la operación y tal y como se le prometió todo va bien, ya que ciertas capacidades han mejorado (quizá tenga

mejor memoria). Entonces, ¿sigue usted siendo usted? Ciertamente sus amigos así lo creen. Y usted también lo cree. No hay ninguna buena razón para decir que de repente usted es una persona diferente. Obviamente, usted ha sufrido la operación para cambiar algo, pero sigue siendo el mismo. Su identidad no ha cambiado. La conciencia de otra persona no ha invadido su cuerpo de repente.

Entonces, motivado por estos resultados, usted decide pasar por otra operación, que esta vez tiene que ver con una región del cerebro diferente. El resultado es el mismo: usted experimenta cierta mejora en sus capacidades, pero sigue siendo usted mismo.

¿Está claro hasta dónde quiero llegar? Usted continúa sometiéndose a más y más operaciones, su confianza en el procedimiento sigue aumentando, hasta que al final le han cambiado todas las partes de su cerebro. En todos los casos las operaciones fueron hechas con mucho cuidado para preservar todos sus patrones y conexiones neocorticales, de manera que no sufriera ninguna pérdida de personalidad, capacidades o recuerdos. Nunca hubo un usted y un Usted 2, solo ha habido un usted. Nadie, ni siquiera usted, ha notado que usted dejara de existir. De hecho, usted está presente.

¿A qué conclusión llegamos? Usted sigue existiendo. Esto no presenta ningún dilema.

Todo está bien. *Excepto por una cosa*: usted, después de los graduales procesos de recambio, es completamente equivalente al Usted 2 del experimento mental anterior. A esto lo llamo el escenario de escaneado y ubicación*. Después de llegar al escenario de recambio gradual, usted posee todos los patrones y conexiones neocorticales que tenía en un principio, solo que en un sustrato no biológico. Esto también es así en el caso de Usted 2 en el escenario de escaneado y ubicación. Después de llegar al escenario de recambio gradual, usted posee algunas capacidades adicionales y una durabilidad mayor que antes del proceso, pero esto es asimismo cierto en el caso de Usted 2 durante el proceso de escaneado y ubicación.

Sin embargo, la conclusión a la que llegamos fue que Usted 2 *no* es usted. Y si, después del proceso de recambio gradual, usted es completamente equivalente a Usted 2 tras el proceso de escaneado y ubicación, entonces tras el proceso de recambio gradual usted tampoco debería ser usted mismo.

Sin embargo, esto contradice nuestra conclusión anterior. El proceso de recambio gradual consta de muchos pasos. Todos esos pasos parecen preservar la identidad, tal y como pensamos hoy en día que los pacientes de Párkinson poseen la misma identidad después de haberles sido instalado un implante neuronal (22).

Este tipo de dilemas filosóficos lleva a algunas personas a la conclusión de que estos escenarios de recambio nunca se producirán (aunque ya están teniendo lugar). Considérese lo siguiente: lo cierto es que todos nosotros sufrimos un proceso de recambio gradual durante la vida. La mayoría de las células de nuestro cuerpo están

siendo remplazadas continuamente. (Mientras leía la última frase usted ha recambiado 100 millones de ellas). Las células de la pared interior del intestino delgado son regeneradas más o menos cada semana, al igual que lo hace la pared protectora del estómago. La esperanza de vida de las células sanguíneas va desde unos pocos días hasta unos pocos meses, dependiendo de su tipo. Las plaquetas duran unos nueve días.

Las neuronas perduran, pero sus orgánulos y moléculas constituyentes son regenerados mensualmente (23). La vida media del microtúbulo de una neurona es de más o menos 10 minutos; la de los filamentos de actina de las dendritas es de unos 40 segundos; las proteínas que proporcionan energía a las sinapsis son remplazadas cada hora; los receptores NMDA de las sinapsis tienen una vida relativamente larga de 5 días.

De manera que usted, en cuestión de meses, es completamente recambiado, lo cual es comparable al escenario de recambio gradual que he descrito más arriba. ¿Es usted la misma persona que era hace unos meses? Ciertamente existen algunas diferencias. Quizá haya aprendido unas cuantas cosas. Sin embargo, usted da por hecho que su identidad persiste, ya que usted no es constantemente destruido y reconstruido.

Consideremos el caso de un río como el que pasa delante de mi oficina. Al mirar ahora a lo que la gente llama Charles River, ¿se trata del mismo río que vi ayer? Reflexionemos primero sobre lo que es un río. El diccionario lo define como “una corriente natural y grande de agua en movimiento”. Según esta definición, el río al que estoy mirando es completamente diferente del río de ayer. Todas y cada una de sus moléculas de agua ha cambiado en un proceso que ocurre muy rápidamente. Así, el filósofo griego Diógenes Laercio escribió en el siglo III a. C. que “no podemos bañarnos dos veces en el mismo río”.

Sin embargo, no es así como solemos pensar sobre los ríos. A la gente le gusta mirarlos porque son símbolos de continuidad y estabilidad. Según la opinión general, el Río Charles al que miré ayer es el mismo río que veo hoy. Nuestras vidas se parecen mucho a esto. En lo fundamental no somos las cosas que constituyen nuestros cuerpos y cerebros. Dichas partículas básicamente fluyen a través de nosotros de la misma forma que las moléculas de agua fluyen a través del río. Somos un patrón que cambia lentamente pero que posee estabilidad y continuidad, aunque las cosas que constituyen el patrón cambian rápidamente.

La introducción gradual de sistemas no biológicos en el interior de nuestros cuerpos y cerebros tan solo será otro ejemplo de la continua regeneración de las partes que nos componen. No alterará la continuidad de nuestra identidad más de lo que lo hace el recambio natural de nuestras células biológicas. Ya hemos externalizado ampliamente nuestros recuerdos históricos, intelectuales, sociales y

personales en nuestros dispositivos y en la nube. Puede que los dispositivos con los que interactuamos para tener acceso a estos recuerdos no estén todavía en el interior de nuestros cuerpos y cerebros, pero a medida que se vuelvan más y más pequeños (y estamos dividiendo el volumen en tres dimensiones de la tecnología por 100 más o menos cada década), lo irán logrando. En cualquier caso, será un buen sitio para dejarlos porque así no los perderemos. Si la gente elige no colocar dispositivos microscópicos en el interior de sus cuerpos no pasará nada, ya que habrá otras maneras de acceder a la ubicua inteligencia de la nube.

Sin embargo, nos volvemos a encontrar en el dilema que he expuesto antes. Usted, después de un periodo de recambio gradual, es equivalente a Usted 2 en un escenario de escaneado y ubicación. No obstante, hemos llegado a la conclusión de que en dicho escenario Usted 2 no posee la misma identidad que tiene usted. Entonces, ¿en qué lugar quedamos nosotros?

Esto nos lleva a reconocer una capacidad que tienen los sistemas no biológicos y que los sistemas biológicos no tienen: la capacidad de ser copiado, guardado en una copia de seguridad y ser reconstruido. Esto es lo que hacemos de forma rutinaria con nuestros dispositivos. Cuando nos hacemos con un smartphone nuevo copiamos en él todos nuestros ficheros, de manera que posee prácticamente la misma personalidad, capacidad y recuerdos que el antiguo smartphone. Quizás también tenga algunas capacidades nuevas, pero el contenido del teléfono antiguo siguen con nosotros. De forma similar, un programa como Watson ciertamente posee copias de seguridad. Si mañana el hardware de Watson fuera destruido, Watson podría ser fácilmente reconstruido a partir de sus copias de seguridad guardadas en la nube.

Esto representa una capacidad del mundo no biológico que no existe en el mundo biológico. Se trata de una ventaja, no de una limitación, y entre otras cosas por eso somos tan dados a día de hoy a cargar nuestros recuerdos en la nube. Seguro que continuaremos avanzando en esta dirección a medida que los sistemas no biológicos adquieran más y más capacidades similares a las de nuestros cerebros biológicos.

He aquí como yo resuelvo el dilema: no es cierto que Usted 2 no sea usted, *sí que es usted*. Lo único que pasa es que ahora hay dos ustedes. Esto no es tan malo. Si cree que usted es algo bueno, entonces dos mejor que uno.

Lo que yo creo que pasará en realidad es que continuaremos por el camino del recambio gradual y del escenario de mejora hasta que en último término la mayor parte de nuestro pensamiento se encuentre en la nube. Mi acto de fe en cuanto a la identidad es que esta se preserva a través de la continuidad del patrón de información que nos hace ser quien somos. La continuidad permite el cambio continuo, de manera que aunque soy algo diferente a quien era ayer, sigo teniendo la misma identidad. Sin embargo, la continuidad del patrón que constituye mi identidad no depende del sustrato. Los sustratos biológicos son maravillosos y nos han llevado muy lejos, sin embargo tenemos muy buenas razones para estar creando un sustrato

más capaz y duradero.

* La mujer es el neurocientífico que se ha citado anteriormente. El autor realiza a menudo este juego de géneros.
* *The busy beaver problem.*
* Chiste sobre el famoso lema cartesiano cogito ergo sum, “pienso, luego existo” (en inglés: “I think, therefore I am”). Descartes, al decir pienso que no (“I think not”) deja de existir ☺.
* “Free won’t” en el original (en contraposición a “free will”, libre albedrío).
* “Scan-and-instantiate scenario” en el original.

CAPÍTULO DIEZ

# La ley de los rendimientos acelerados aplicada al cerebro

Si bien, en algunos aspectos, el hombre debería seguir siendo la criatura más elevada, ¿no concuerda esto con las maneras de la naturaleza, las cuales permiten que animales que en general han sido ampliamente superados mantengan su superioridad en ciertas cosas? ¿No es cierto que [la naturaleza] le ha permitido a la hormiga y a la abeja mantener su superioridad sobre el hombre en lo que se refiere a la organización de sus comunidades y de sus disposiciones sociales, al pájaro en lo que se refiere a atravesar el aire, al pez en el nadar, al caballo en la fuerza y la fugacidad, y al perro en el sacrificio?

—SAMUEL BUTLER, 1871

Hubo un tiempo en el que La Tierra carecía completamente de vida tanto animal como vegetal y, según nuestros mejores filósofos, era una mera pelota caliente cuya corteza se iba enfriando gradualmente. Si un ser humano hubiera existido durante el periodo en el que La Tierra se encontraba en este estado y se le hubiera permitido contemplarla como si fuera otro mundo con el que él no guardaba relación, y si al mismo tiempo ignorara por completo la física, ¿no habría considerado imposible que criaturas dotadas de cualquier tipo de consciencia pudieran evolucionar a partir de las cenizas que observaría? ¿No habría rechazado cualquier posibilidad de que tuviera el potencial de albergar consciencia? Y sin embargo, con el transcurrir del tiempo, surgió la consciencia. ¿No es posible entonces que se den otros canales nuevos para la consciencia, pero cuyos signos en el presente no podamos detectar?

—SAMUEL BUTLER, 1871

Si reflexionamos sobre las diferentes fases de la vida y de la consciencia a las que la evolución ya ha dado lugar, sería imprudente decir que es imposible que evolucionen otras y que la vida animal es el final de todas las cosas. Hubo un tiempo en que el fuego era el final de todo, y otro en el que las piedras y el agua también lo fueron.

—SAMUEL BUTLER, 1871

Basándonos solamente en la poca consciencia que poseen las máquinas hoy en día, no se puede asegurar que en último término sea imposible el desarrollo de conciencia mecánica. Un molusco tampoco posee mucha consciencia. Piénsese sobre el extraordinario avance que han hecho las máquinas durante los últimos siglos y fijémonos en lo lento que avanzan los reinos animal y vegetal. Por decirlo de alguna manera, las máquinas más complejas no son tanto criaturas de ayer, como de los últimos cinco minutos, si las comparamos con épocas pasadas. Imaginémonos por un momento que los seres conscientes hubieran existido desde hace 20 millones de años, ¡qué grandes avances han experimentado las máquinas en el último milenio! ¿No es verdad que el mundo podría perdurar otros 20 millones de años más? Entonces ¿qué será en lo que no acabarán convirtiéndose?

—SAMUEL BUTLER, 1871

Mi tesis principal, a la que doy el nombre de ley de los rendimientos acelerados (LOAR)*, es que las mediciones fundamentales en el campo de la tecnología de la información siguen trayectorias predecibles y exponenciales, lo cual va en contra de la sabiduría popular que mantiene que "el futuro no se puede predecir". Ciertamente hay muchas cosas (qué proyecto, empresa o estándar tecnológico dominará el mercado, cuándo habrá paz en oriente medio) que son impredecibles. Sin embargo, el ratio entre rendimiento y precio en relación a las capacidades de la información ha demostrado ser extraordinariamente predecible. Así, es sorprendente observar cómo estas tendencias no se ven perturbadas por circunstancias tales como la guerra o la paz, la prosperidad o la recesión.

Una de las razones principales por las cuales la evolución dio lugar a los cerebros fue la predicción del futuro. Cuando, hace miles de años, nuestros ancestros caminaban por las sabanas y se daban cuenta de que se iban a cruzar con un animal que se dirigía en la misma dirección en la que ellos iban, predecían que si seguían caminando sus caminos se encontrarían. Por lo tanto decidían cambiar de dirección. Así se demuestra que la previsión es valiosa para asegurar la supervivencia.

Sin embargo, estos pronosticadores de serie son lineales, no exponenciales, siendo esta linealidad una cualidad que proviene de la organización asimismo lineal del neocórtex. Recuérdese que el neocórtex está constantemente haciendo predicciones (qué letra o palabra veremos a continuación, a quién esperamos encontrarnos al doblar la esquina, etc). El neocórtex está organizado según secuencias lineales de pasos que forman parte de patrones, lo cual significa que el pensamiento exponencial no es algo que se dé en nosotros de forma natural. El cerebelo también utiliza predicciones lineales. Al ayudarnos a agarrar al vuelo una pelota, lo que realiza es una predicción lineal sobre la situación que tomará la pelota dentro de nuestro campo visual y sobre dónde deberíamos colocar la mano dentro de nuestro campo visual para agarrarla.

Tal y como he expuesto, existe una gran diferencia entre las progresiones lineales y exponenciales (linealmente, 40 pasos son 40 pasos, pero exponencialmente son un billón). Por eso mis predicciones basadas en la ley de los rendimientos acelerados pueden parecer en un principio un tanto sorprendentes, ya que tenemos que entrenarnos a nosotros mismos para pensar exponencialmente. Esta es la forma correcta de pensar en lo que se refiere a las tecnologías de la información.

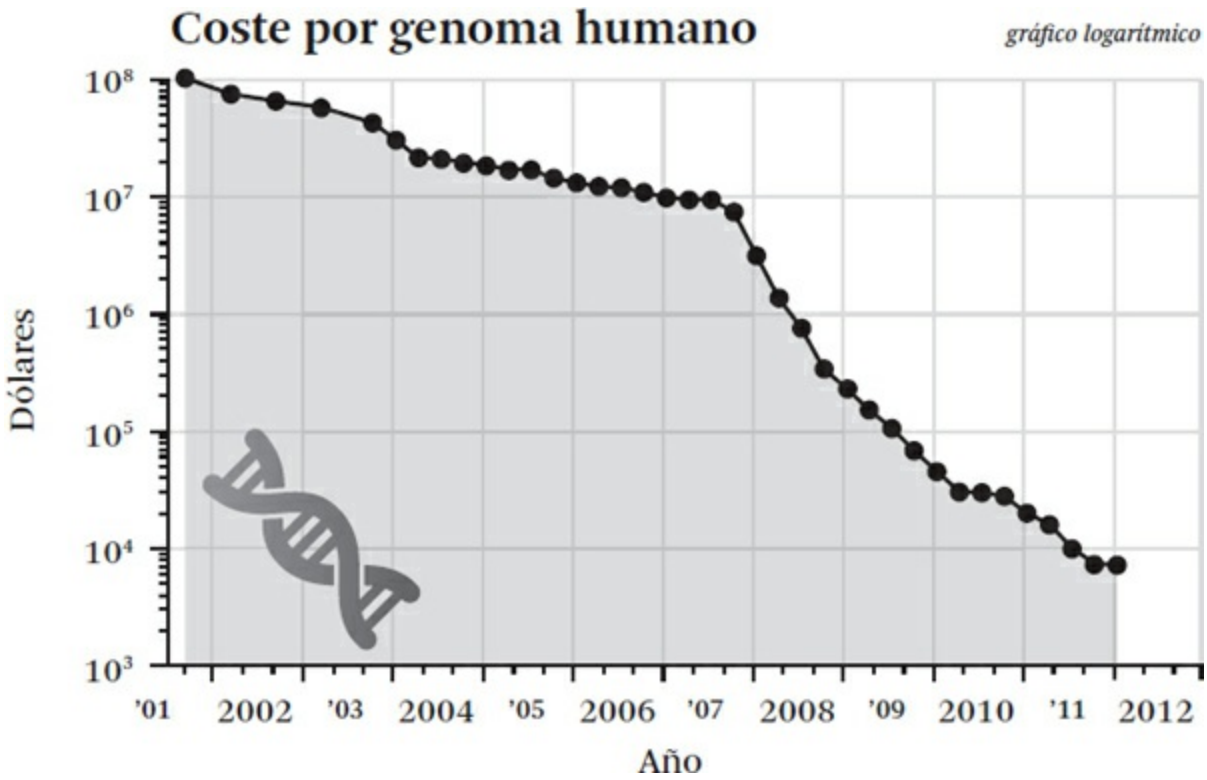
El ejemplo paradigmático de la ley de los rendimientos acelerados es el crecimiento constante y doblemente exponencial en la relación entre rendimiento y precio en el campo de la informática, el cual se ha mantenido durante 110 años pese a dos guerras mundiales, la gran depresión, la guerra fría, el colapso de la Unión Soviética, el resurgimiento de China, la última crisis financiera y todos los hechos destacables de finales del siglo XIX, del siglo XX y de principios del siglo XXI.

Algunas personas se refieren a este fenómeno como “ley de Moore“, pero esto es producto de una confusión. La ley de Moore, que sostiene que cada dos años se pueden colocar el doble de componentes en un circuito integrado haciendo que estos funcionen más deprisa dado que son más pequeños, solo es un paradigma entre muchos. De hecho fue el quinto, no el primero de los paradigmas que condujo a un crecimiento exponencial en la relación entre rendimiento y precio en el campo de la informática.

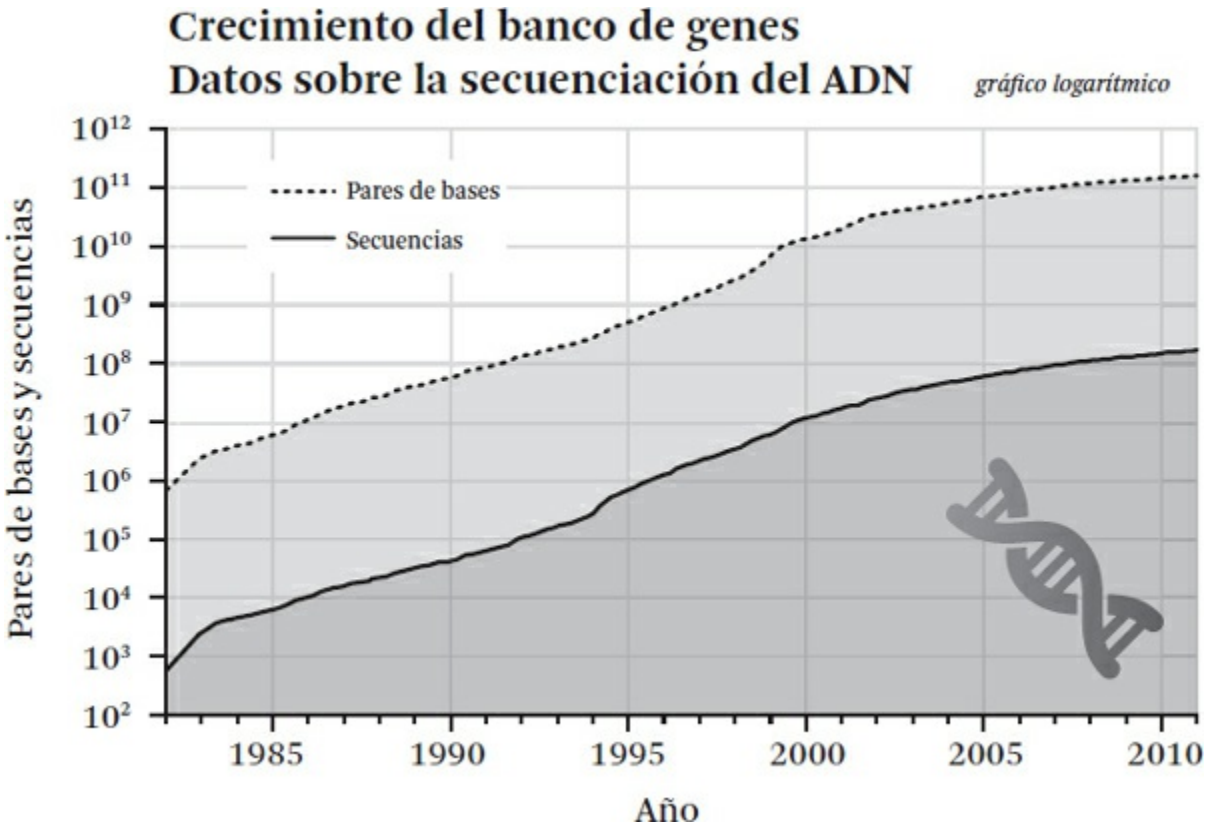
El auge exponencial de la informática empezó con el censo que se realizó en EE.UU. en el año 1890. Décadas antes de que Gordon Moore naciera, este censo fue el primero en hacerse con medios automatizados y utilizando el primer paradigma del cálculo electromecánico. En *La Singularidad está cerca* proporcioné este mismo gráfico hasta el año 2002 y aquí lo actualizo hasta el año 2009 (véase el gráfico de la página 243 titulado “crecimiento exponencial de la computación durante 110 años). Como verá, la constante y predecible trayectoria de la curva ha continuado pese a la reciente desaceleración económica.

La capacidad de cálculo es el ejemplo más importante de la ley de los rendimientos acelerados debido a la gran cantidad de datos que disponemos sobre ella, a la ubicuidad de la computación y a su fundamental papel en la revolución que ha acabado transformando todo aquello que consideramos importante. Sin embargo, está lejos de ser el único ejemplo. Una vez que una tecnología se convierte en una tecnología de la información, esta pasa a estar sujeta a ley de los rendimientos acelerados.

El nuevo área de la biomedicina se está convirtiendo en el ejemplo más importante de una tecnología y una industria que están siendo transformadas de esta manera. Históricamente, el progreso en la medicina se ha basado en descubrimientos accidentales. Así, el progreso en épocas pasadas fue lineal, no exponencial. Sin embargo, esto ha resultado beneficioso. La esperanza de vida ha aumentado desde los 23 años de hace un milenio, pasando por los 37 años de hace dos siglos, hasta llegar a los casi 80 años de hoy. Así, mediante la comprensión del software de la vida (el genoma), la medicina y la biología se han convertido en tecnologías de la información. El propio proyecto del genoma humano fue completamente exponencial, ya que anualmente la cantidad de datos genéticos se dobló y el coste por par de bases se dividió por dos desde que el proyecto se iniciara en 1990 (3). (Todos los gráficos de este capítulo han sido actualizados desde la publicación de *La Singularidad está cerca*).




**El coste de secuenciar un genoma de tamaño humano (1).**



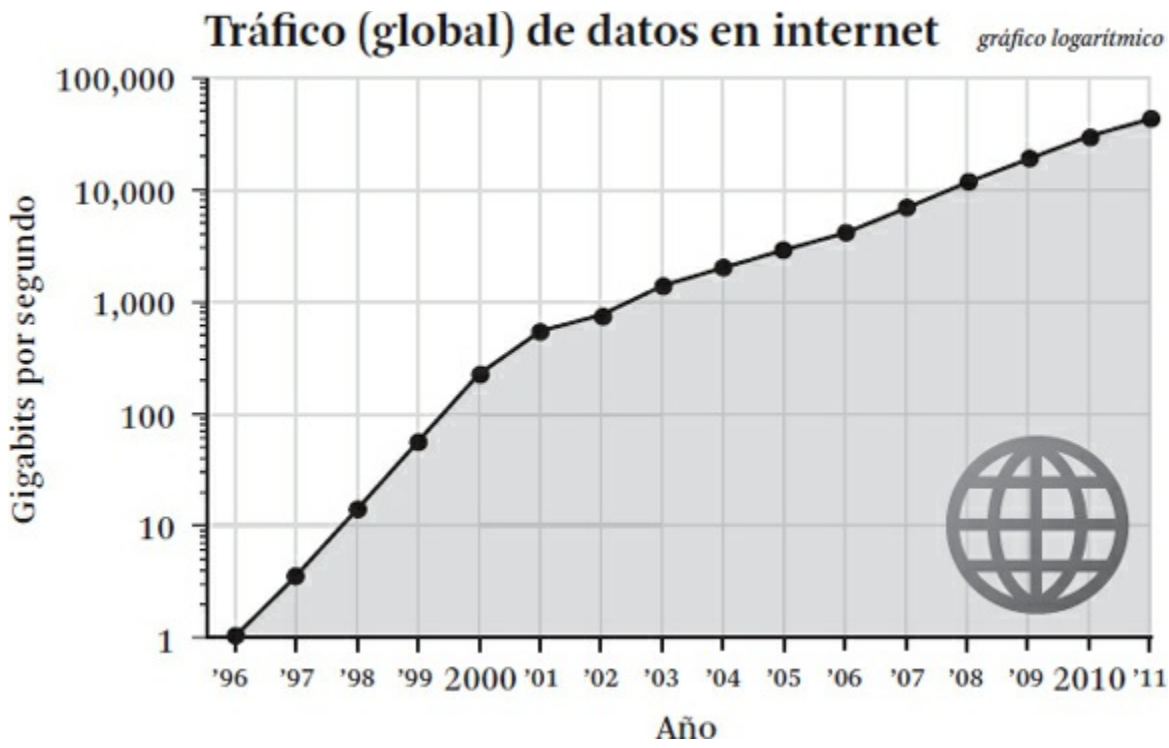

**La cantidad de datos genéticos secuenciados en el mundo anualmente (2).**

Ahora poseemos la capacidad para diseñar intervenciones biomédicas sobre

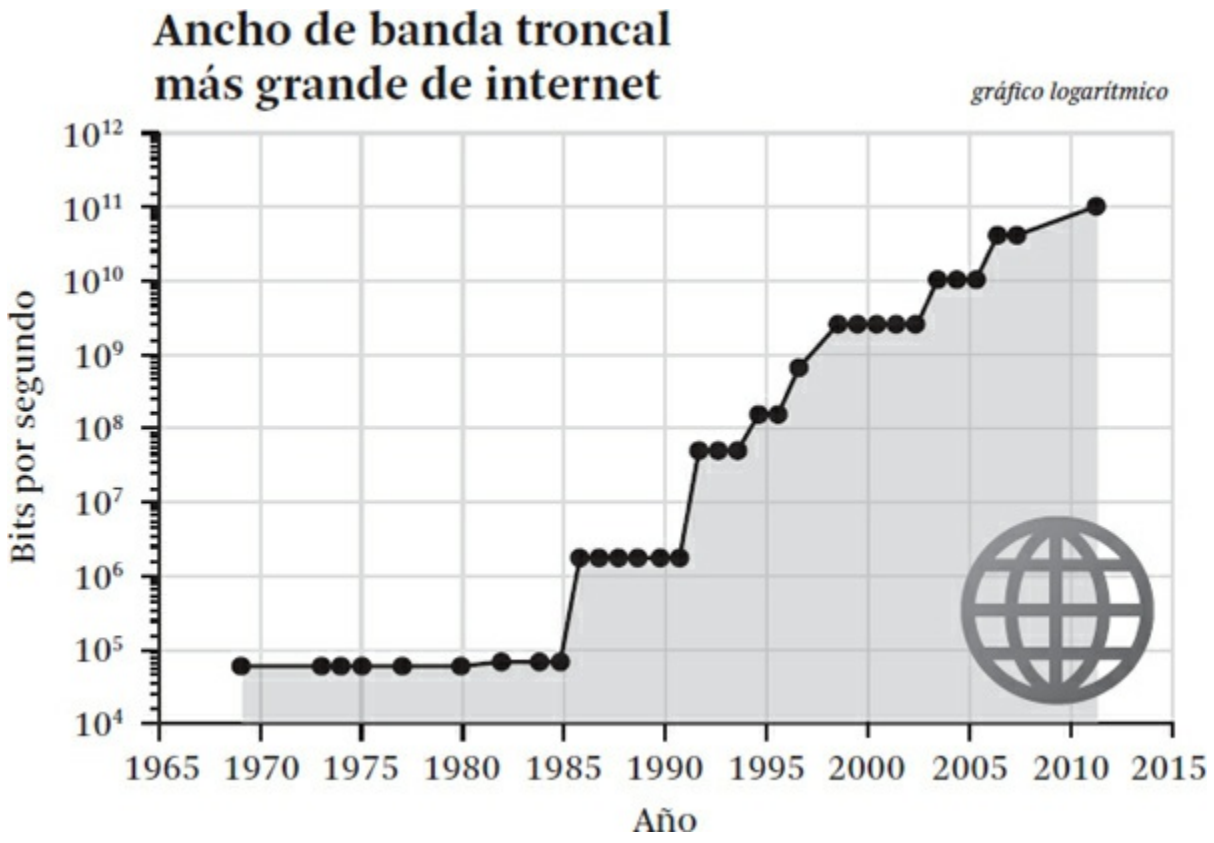
ordenadores y probarlas en simuladores biológicos (tanto la escala como la precisión de estos procedimientos también se dobla anualmente). También podemos actualizar nuestro obsoleto software. El ARN interferente puede desactivar genes y nuevas formas de terapias génicas pueden añadir nuevos genes, no solo a un recién nacido, sino también a un individuo maduro. El avance de las tecnologías génicas afecta asimismo al proyecto de aplicar la ingeniería inversa al cerebro, ya que un importante aspecto de este proyecto es comprender cómo los genes controlan funciones cerebrales tales como la creación de nuevas conexiones que reflejen conocimientos corticales recientemente adquiridos. Existen muchas otras manifestaciones de esta integración entre biología y tecnología de la información que van apareciendo a medida que nos adentramos más allá de la secuenciación del genoma y entramos en el campo de la síntesis del mismo.

Otra tecnología de la información que ha mostrado un crecimiento exponencial constante es nuestra capacidad para comunicarnos los unos con los otros y la transmisión de enormes depósitos de conocimientos humanos. Existen muchas maneras de medir este fenómeno. La ley de Cooper, que sostiene que la capacidad total de bits de las comunicaciones inalámbricas en un determinado espectro radioeléctrico se dobla cada 30 meses, se ha cumplido desde la época en que Guglielmo Marconi utilizó el telégrafo inalámbrico para enviar transmisiones en código Morse en 1897 hasta las tecnologías de la comunicación en 4G que usamos hoy en día (4). Según la ley de Cooper, la cantidad de información que puede ser transmitida a través de un espectro radioeléctrico determinado se ha venido doblando anualmente cada dos años y medio desde hace más de un siglo. Otro ejemplo es el número de bits por segundo transmitidos a través de internet, que se está doblando cada año y cuarto (5).

La razón por la cual me empezó a interesar la predicción de ciertos aspectos de la tecnología es que hará unos 30 años me di cuenta de que la clave para convertirme en un inventor de éxito (una profesión por la que me decidí cuando tenía 5 años) era e l *timing*. La mayoría de las invenciones e inventores fracasan no porque sus dispositivos no funcionen, sino porque su *timing* es incorrecto y los dispositivos aparecen o bien antes de que todos los factores propicios estén en su lugar o bien demasiado tarde, por lo que dejan escapar sus oportunidades.




**El ancho de banda internacional (país a país) que dedica el mundo a internet (6).**




**El ancho de banda (velocidad) troncal más grande de internet (7).**

En calidad de ingeniero, hará unas tres décadas que empecé a recopilar datos sobre las mediciones de la tecnología en diferentes áreas. Cuando comencé con este

cometido no esperaba encontrarme con una imagen clara, pero sí que esperaba que me proporcionara alguna orientación y que me permitiera hacer suposiciones con sentido. Mi objetivo era (y sigue siendo) acompasar apropiadamente mis creaciones tecnológicas para que se adecúen al mundo en el que existirán cuando su proyecto se complete, ya que entonces me di cuenta de que dicho mundo iba a ser muy diferente del mundo en el que fueron concebidas.

Pensemos en lo mucho y lo rápido que ha cambiado el mundo tan solo en épocas recientes. Hace tan solo unos años la gente no usaba las redes sociales. Por ejemplo, Facebook se fundó en 2004 y a finales de marzo de 2012 tenía 901 millones de usuarios activos cada mes (8). Luego están los wikis, los blogs, los tweets, etc. En la década de 1990 la mayor parte de la gente no usaba ni buscadores ni teléfonos móviles. Imagínese un mundo sin ellos. Parece como si se tratara de historia antigua, pero no hace tanto tiempo, y el mundo cambiará todavía más drásticamente en el futuro cercano.

Durante el curso de mi investigación hice un descubrimiento asombroso: si una tecnología es una tecnología de la información, las mediciones básicas en cuanto a rendimiento/precio y capacidad (por unidad de tiempo, coste o cualquier otro recurso) describirán trayectorias exponenciales asombrosamente precisas.

Además, estas trayectorias dejan atrás los paradigmas específicos en los que se basan (tal y como ocurre con la ley de Moore). Sin embargo, cuando un paradigma pierde su fuelle (por ejemplo, cuando en la década de 1950 los ingenieros dejaron de ser capaces de reducir el tamaño y coste de los tubos de vacío), se crea una presión investigadora para producir el siguiente paradigma, de manera que el progreso comienza otra nueva curva en S.

Entonces, la parte exponencial de dicha curva en S del nuevo paradigma continúa el desarrollo exponencial correspondiente a la medición de la tecnología de la información que estemos analizando. Así, la computación basada en los tubos de vacío que se realizaba en la década de 1950 dejó paso a los transistores en la década de 1960, y luego a los circuitos integrados y a la ley de Moore a finales de la década de 1960 y la época que vino después hasta nuestros días. A su vez, la ley de Moore dejará paso a la computación en tres dimensiones, cuyos primeros ejemplos ya existen. La razón por la que las tecnologías de la información son capaces de transcender permanentemente las limitaciones de cualquier paradigma es que los recursos necesarios para computar, recordar o transmitir un bit de información son cada vez más y más pequeños.

Podemos preguntarnos, ¿existen límites últimos en nuestra capacidad de computar y transmitir información, independientemente de cuál sea el paradigma? La respuesta, basándonos en nuestra comprensión actual de la física computacional, es que sí. Sin embargo, dichos límites no son muy limitantes, ya que, en último término, seremos capaces de expandir nuestra inteligencia multiplicándola billones de veces

gracias a la computación molecular. Según mis cálculos, alcanzaremos estos límites a finales de este siglo.

Es importante señalar que no todo fenómeno exponencial es un ejemplo de la ley de los rendimientos acelerados. Muchos analistas malinterpretan la LOAR al citar tendencias exponenciales que no están basadas en la información. Por ejemplo, señalan que las maquinillas de afeitar para hombres han pasado de tener una cuchilla a tener dos y luego cuatro, y después se preguntan, ¿dónde están las maquinillas de ocho cuchillas? No obstante, las maquinillas de afeitar no forman parte (todavía) de la tecnología de la información.

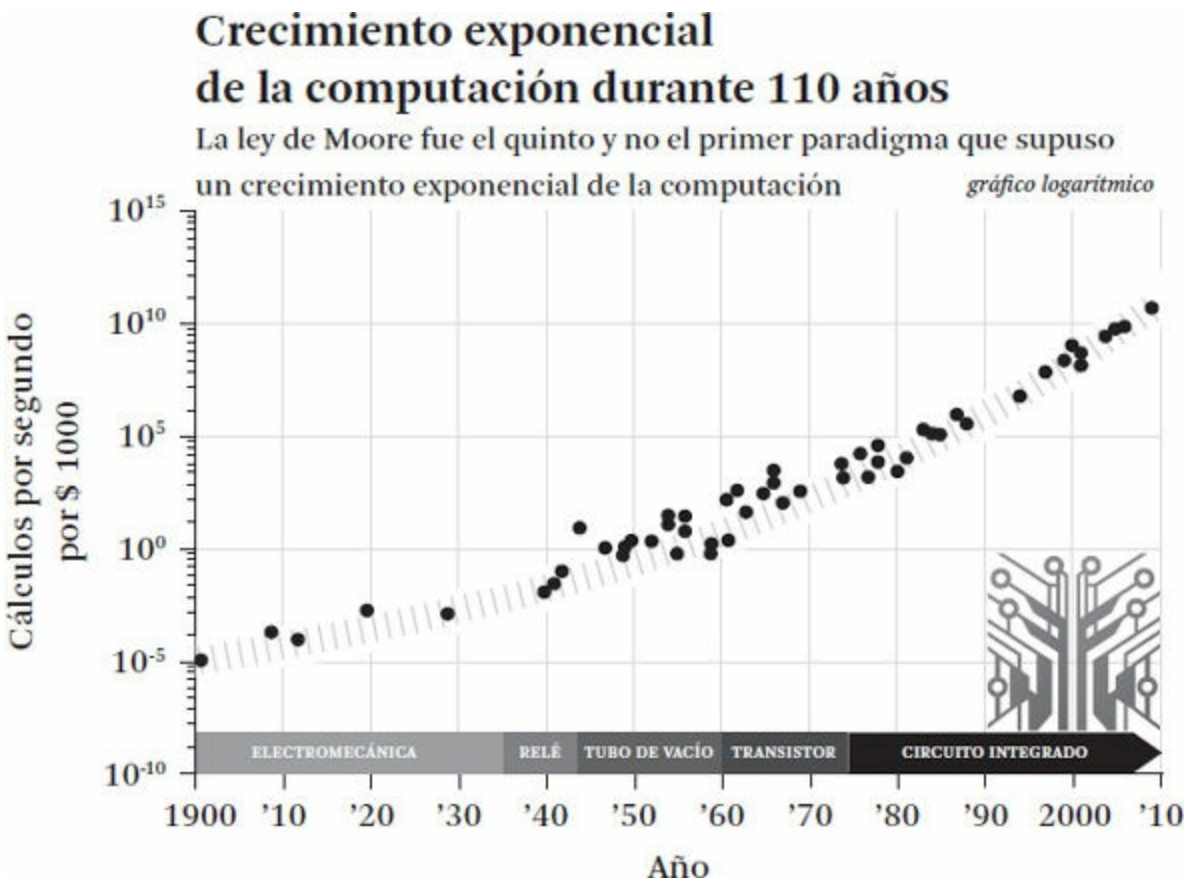
En *La Singularidad está cerca* proporciono un análisis teórico que incluye una explicación matemática en el apéndice del libro para explicar por qué la LOAR es tan extraordinariamente predecible. Esencialmente, siempre utilizamos la última tecnología para crear la siguiente. Las tecnologías se construyen sobre sí mismas de forma exponencial, y este fenómeno es fácilmente medible si involucra una tecnología de la información. En 1990 utilizamos ordenadores y otras herramientas de aquella época para crear los ordenadores de 1991; en 2012 estamos utilizando las herramientas informáticas actuales para crear las máquinas de 2013 y 2014. Para exponerlo de forma más general: esta aceleración y este crecimiento exponencial se cumple en todo proceso en el cual estén involucrados patrones de información. Así, observamos aceleración en el ritmo de la evolución biológica y una aceleración similar (pero mucho más rápida) en la evolución tecnológica, que a su vez es una consecuencia de la evolución biológica.

Empezando con las que hice a mediados de la década de 1980 en mí libro *The Age of Intelligent Machines*, mi historial público ya data de más de un cuarto de siglo de predicciones basadas en la ley de los rendimientos acelerados. Algunos ejemplos de predicciones que se han cumplido y que aparecen en dicho libro son: el nacimiento a mediados y finales de la década de 1990 de una amplia red de comunicaciones llamada *worldwide web* que comunica a personas de todo el mundo las unas con las otras y con todo el conocimiento humano; una gran ola democratizadora que emerge a partir de esta red de comunicación descentralizada y se lleva por delante a la Unión Soviética; la derrota hacia 1998 del campeón del mundo de ajedrez a manos de una máquina; y hay muchas otras.

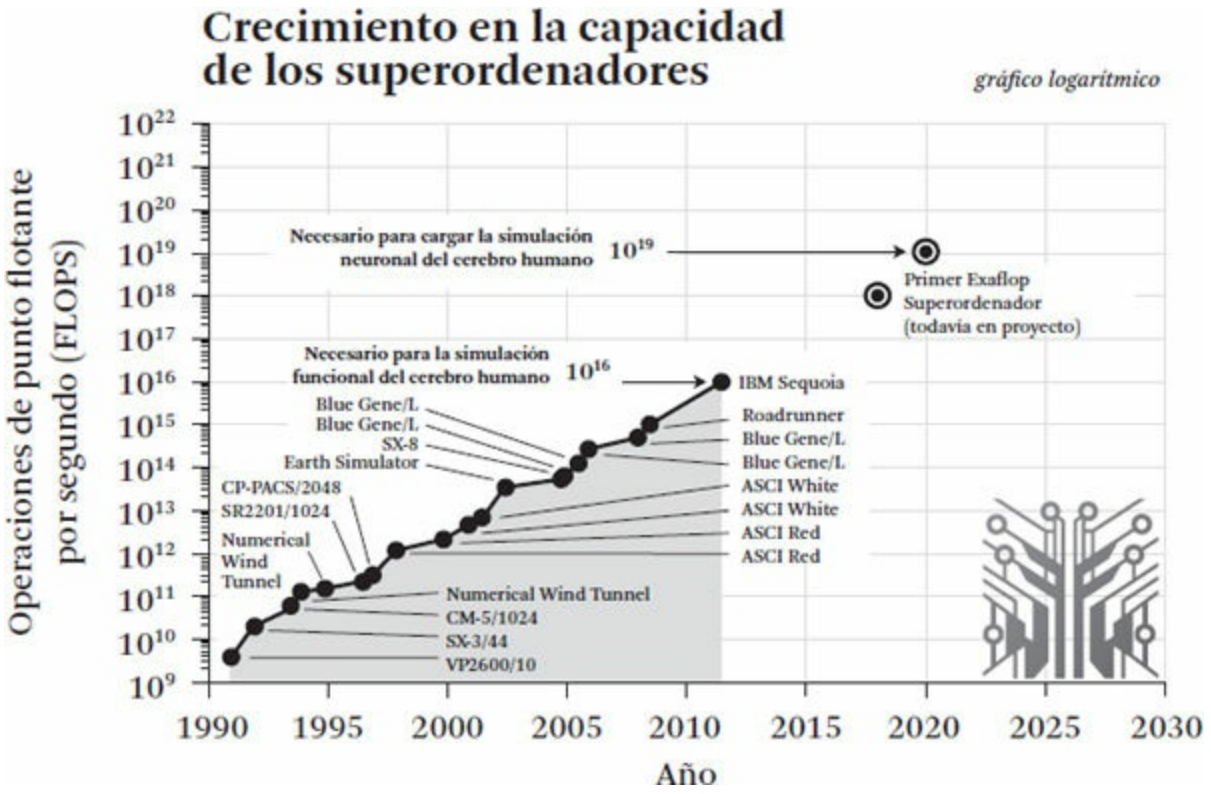
En lo que se refiere a la computación, en mi libro *The Age of Spiritual Machines* describí la ley de los rendimientos acelerados de forma exhaustiva y proporcioné un siglo de datos que mostraban la progresión doblemente exponencial de la relación entre rendimiento y precio de la computación hasta el año 1998 (más abajo estos datos se encuentran actualizados hasta 2009).

Recientemente escribí una reseña de 146 páginas sobre las predicciones que hice en *The Age of Intelligent Machines, The Age of Spiritual Machines* y en *La*

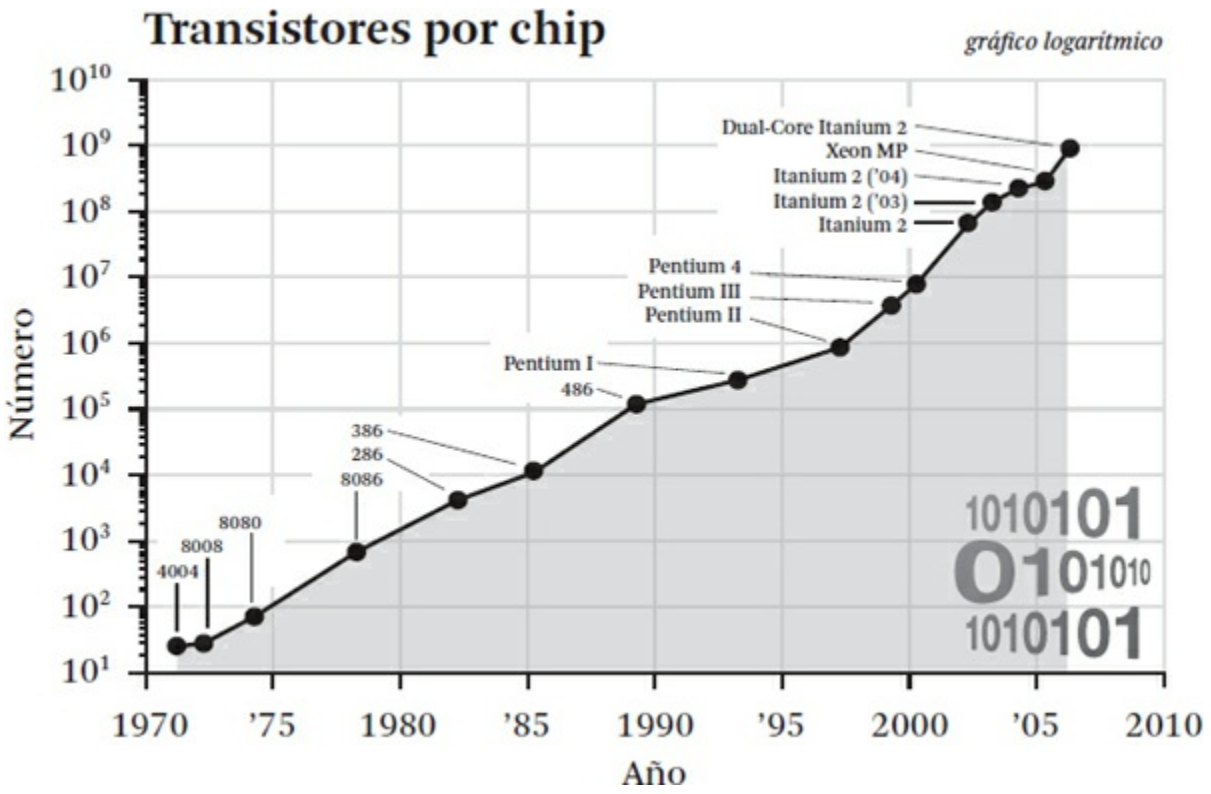
*Singularidad está cerca*. (Este ensayo lo puede leer visitando el link de la nota (9)). *The Age of Spiritual Machines* incluye cientos de predicciones para décadas concretas (2009, 2019, 2029 y 2099). Por ejemplo, en *The Age of Spiritual Machines* (escrito en la década de 1990) hice 147 predicciones para el año 2009. De estas, 115 (el 78%) son completamente correctas a finales de 2009, siendo las predicciones sobre los valores fundamentales de la capacidad y de la relación entre rendimiento y precio de las tecnologías de la información particularmente precisas.



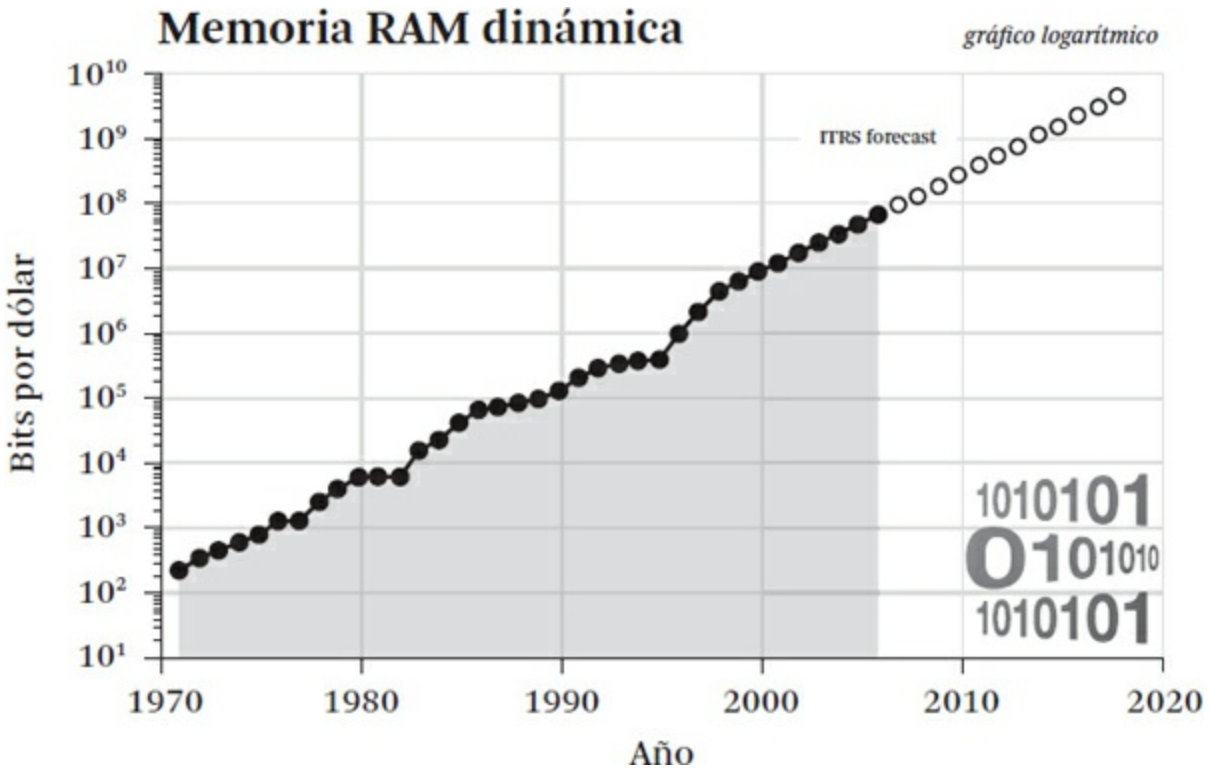

**Cálculos por segundo por 1000 dólares (constantes) en diferentes dispositivos informáticos (10).**



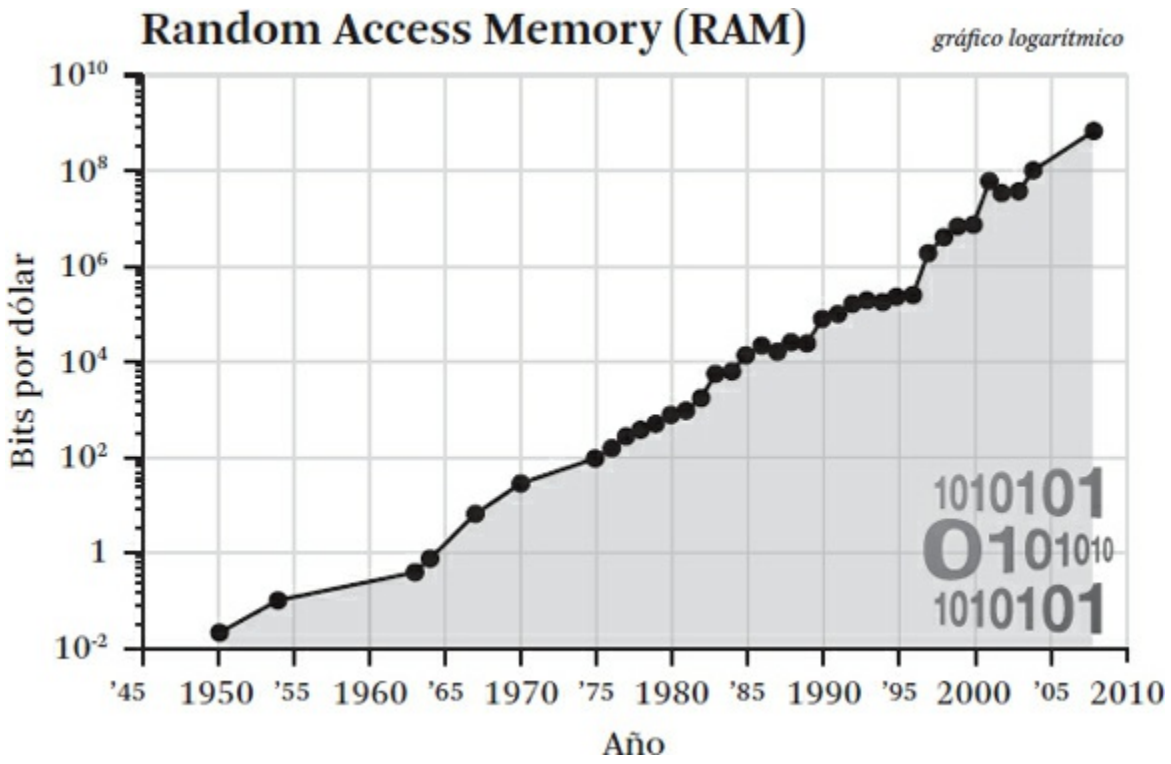

**Operaciones de punto flotante por segundo en diferentes superordenadores (11).**



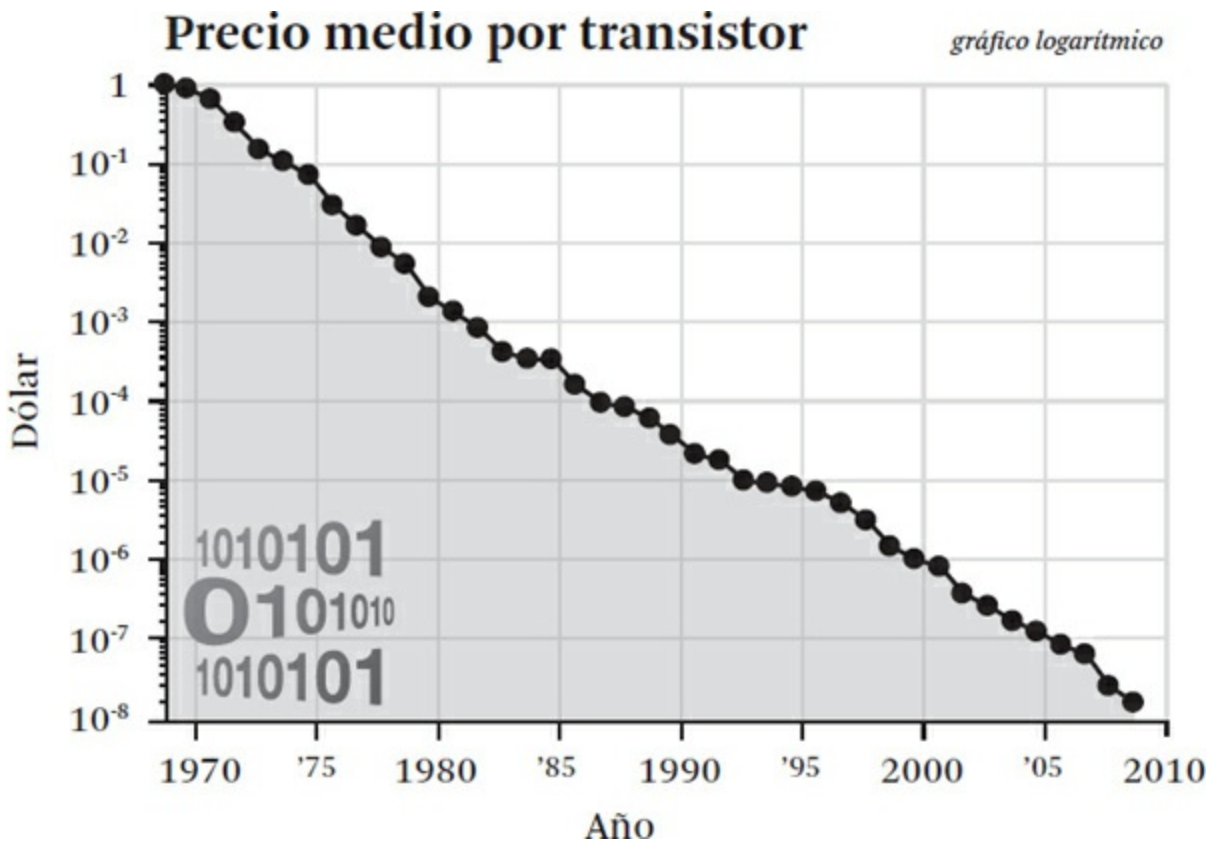

**Transistores por chip en diferentes procesadores Intel (12).**



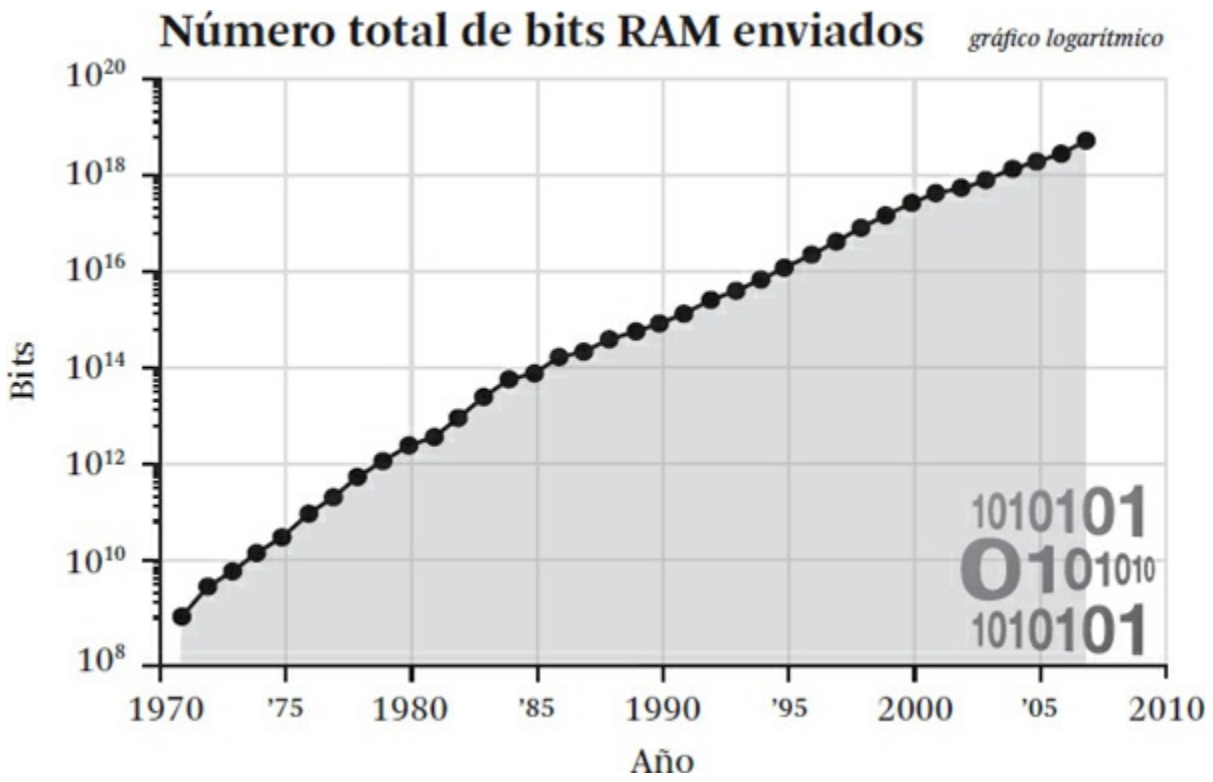

**Bits por dólar en chip de memoria RAM (13).**



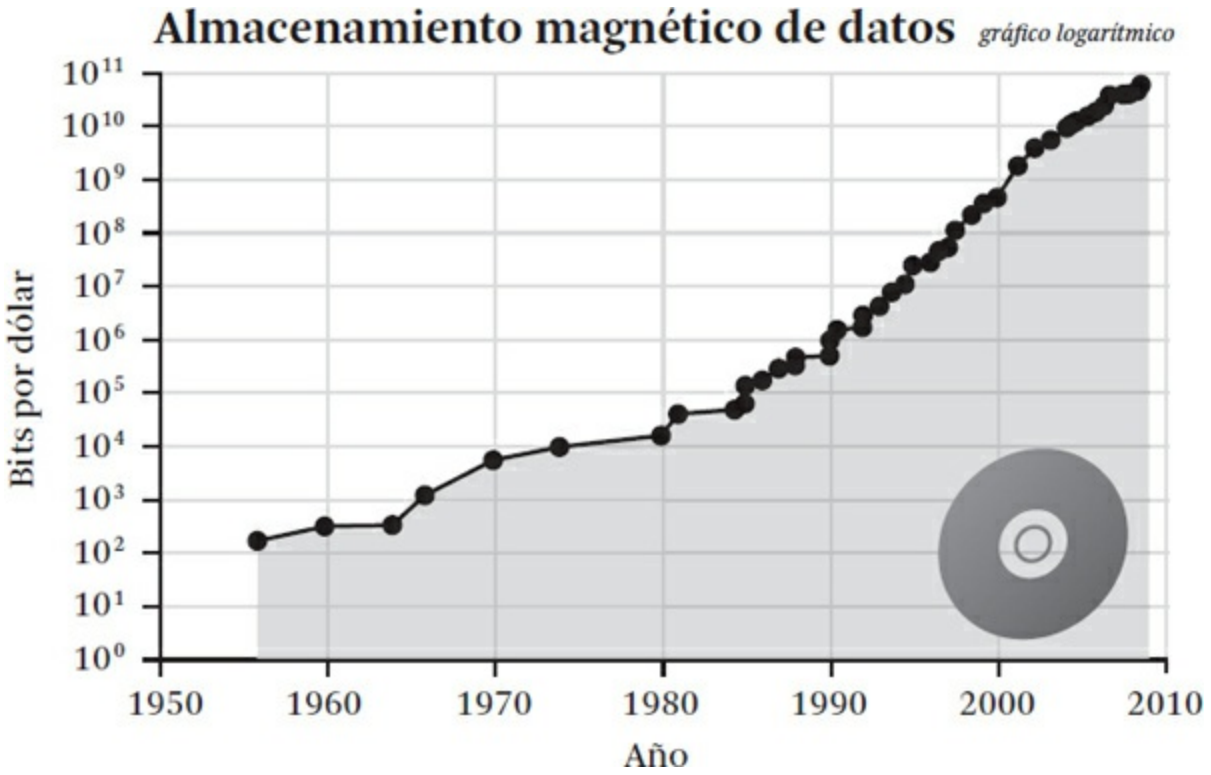

**Bits por dólar en chip de memoria RAM (14).**




**Precio medio por transistor en dólares (15).**




**El número total de bits de memoria RAM enviado cada año (16).**




**Bits por dólar (en dólares constantes del año 2000) para el almacenamiento magnético de datos (17).**

Otras 12 (el 8%) son "esencialmente correctas". Un total de 127 predicciones (el 86%) son correctas o esencialmente correctas. (Como las predicciones eran para una década en concreto, una predicción para 2009 se considera "esencialmente correcta" si se hace realidad en 2010 o 2011). Otras 17 (el 12 %) son parcialmente correctas y 3 (el 2%) están equivocadas.

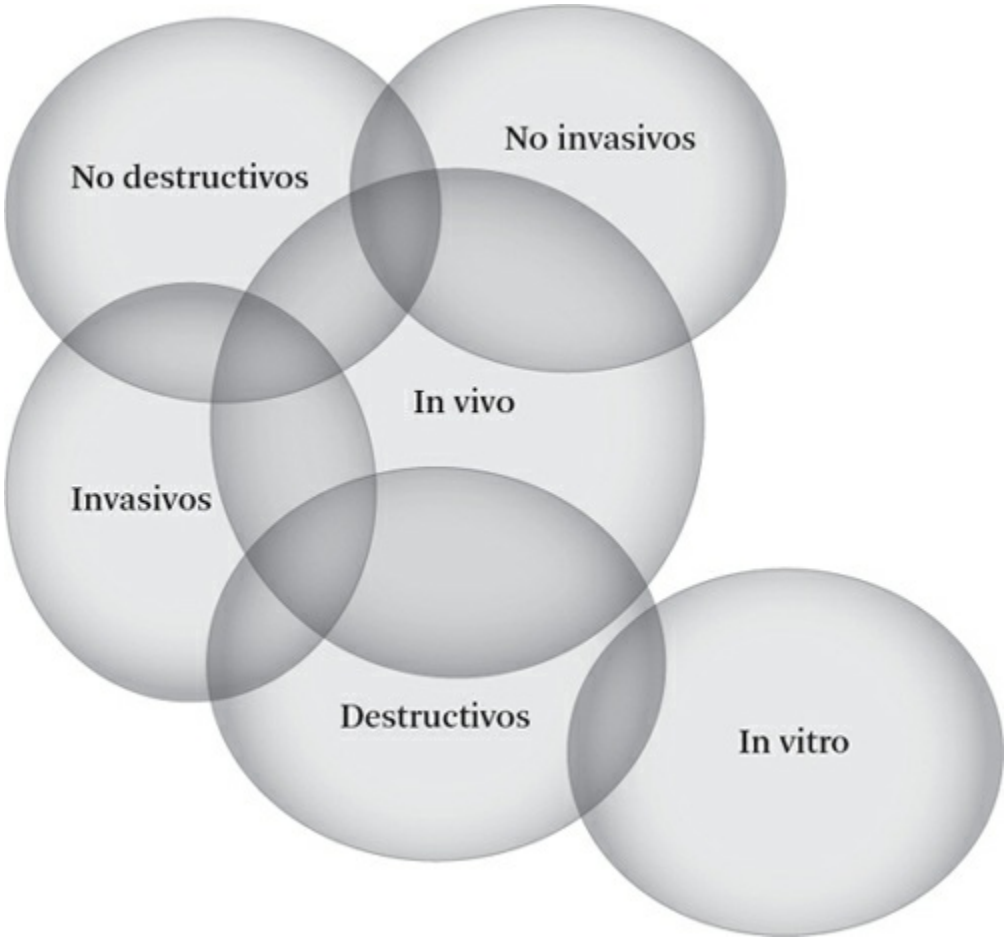
Incluso las predicciones "equivocadas" no son del todo equivocadas. Por ejemplo, considero mi predicción de que tendríamos coches que se conducen solos como equivocada, aunque Google ha exhibido este tipo de coches y en octubre de 2010 cuatro caravanas eléctricas sin conductor completaron con éxito un test de conducción de 13 000 kilómetros desde Italia hasta China (18). Los expertos en la materia predicen ahora que estas tecnologías estarán disponibles para el gran público hacia finales de esta década.

Todas las tecnologías de la computación y de la comunicación se expanden exponencialmente y contribuyen al proyecto de comprender y recrear los métodos del cerebro humano. Este intento no está organizado como un único proyecto, sino que es el resultado de una gran cantidad de proyectos diferentes que incluyen la modelización exhaustiva de los componentes del cerebro (desde neuronas individuales hasta el neocórtex en su totalidad), la cartografía del "connectome" (las conexiones neuronales del cerebro), la simulación de regiones cerebrales y muchas otras cosas. Todo esto ha ido ampliándose exponencialmente. La mayor parte de las

evidencias presentadas en este libro se han hecho públicas recientemente, por ejemplo, el estudio Wedeen de 2012 expuesto en el capítulo 4 que demuestra el muy ordenado y "simple" (para citar a los investigadores) patrón en forma de red que siguen las conexiones del neocórtex. En dicho estudio, los investigadores reconocieron que sus perspectivas e imágenes solo fueron posibles como resultado de la nueva tecnología de imágenes en alta resolución.

Las tecnologías para el escaneo del cerebro están mejorando su resolución, tanto espacial como temporal, a un ritmo exponencial. Los diferentes tipos de métodos de escaneo del cerebro que están siendo probados van desde métodos completamente no invasivos que pueden ser usados en humanos hasta métodos más invasivos o destructivos usados en animales.

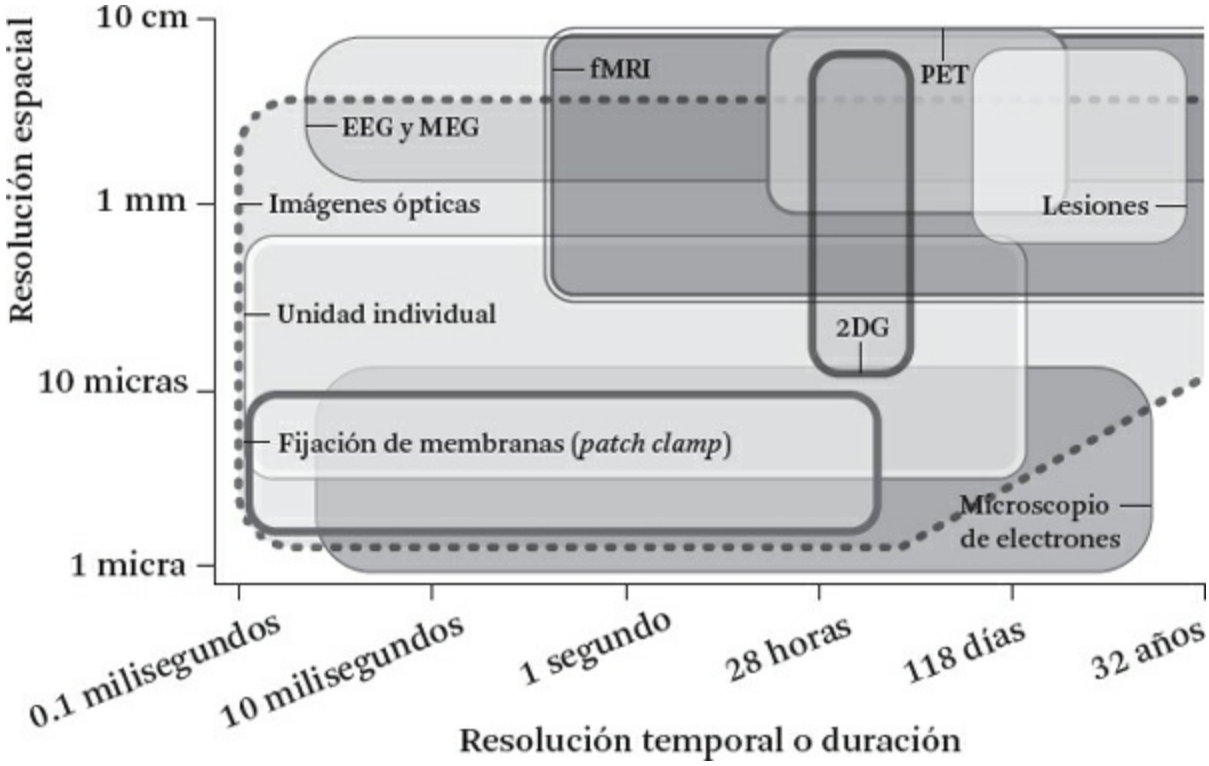
La imagen por resonancia magnética (IRM), una técnica por imágenes no invasiva con una resolución temporal relativamente alta, ha sufrido una constante mejora a ritmo exponencial, hasta el punto de que las resoluciones espaciales se acercan ya a las 100 micras (las millonésimas partes de un metro).



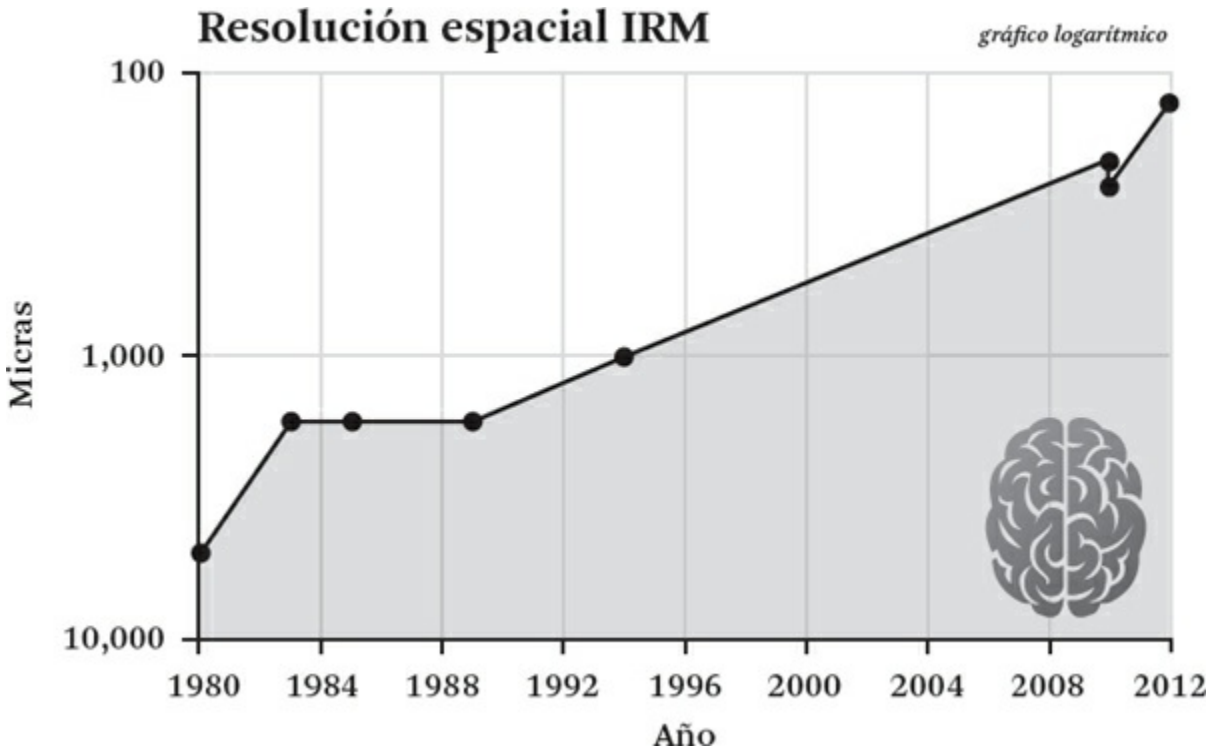

**Un diagrama de Venn sobre los métodos por imágenes del cerebro (19).**

La toma destructiva de imágenes, que se realiza para reunir el connectome (el mapa de todas las conexiones interneuronales) de los cerebros de los animales, también ha mejorado a un ritmo exponencial. Actualmente, la máxima resolución

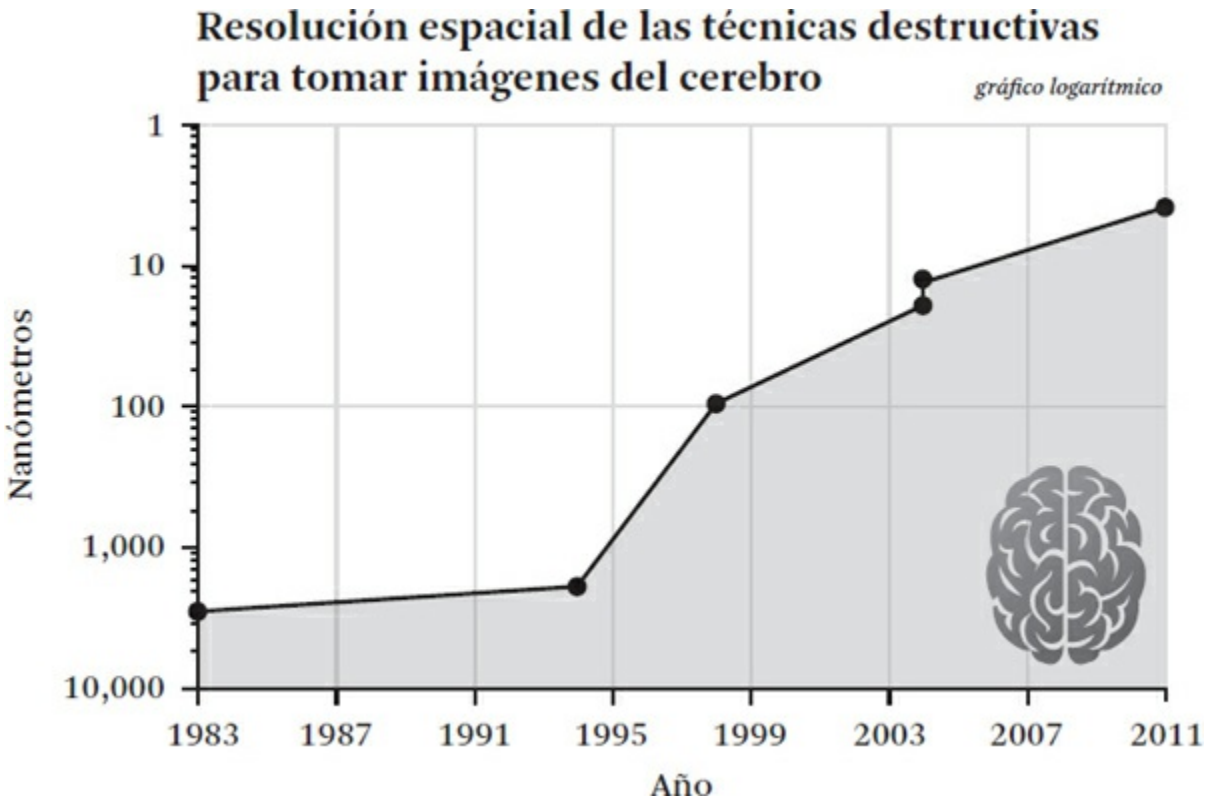
alcanzada es de unos 4 nanómetros, que es suficiente para observar conexiones individuales.



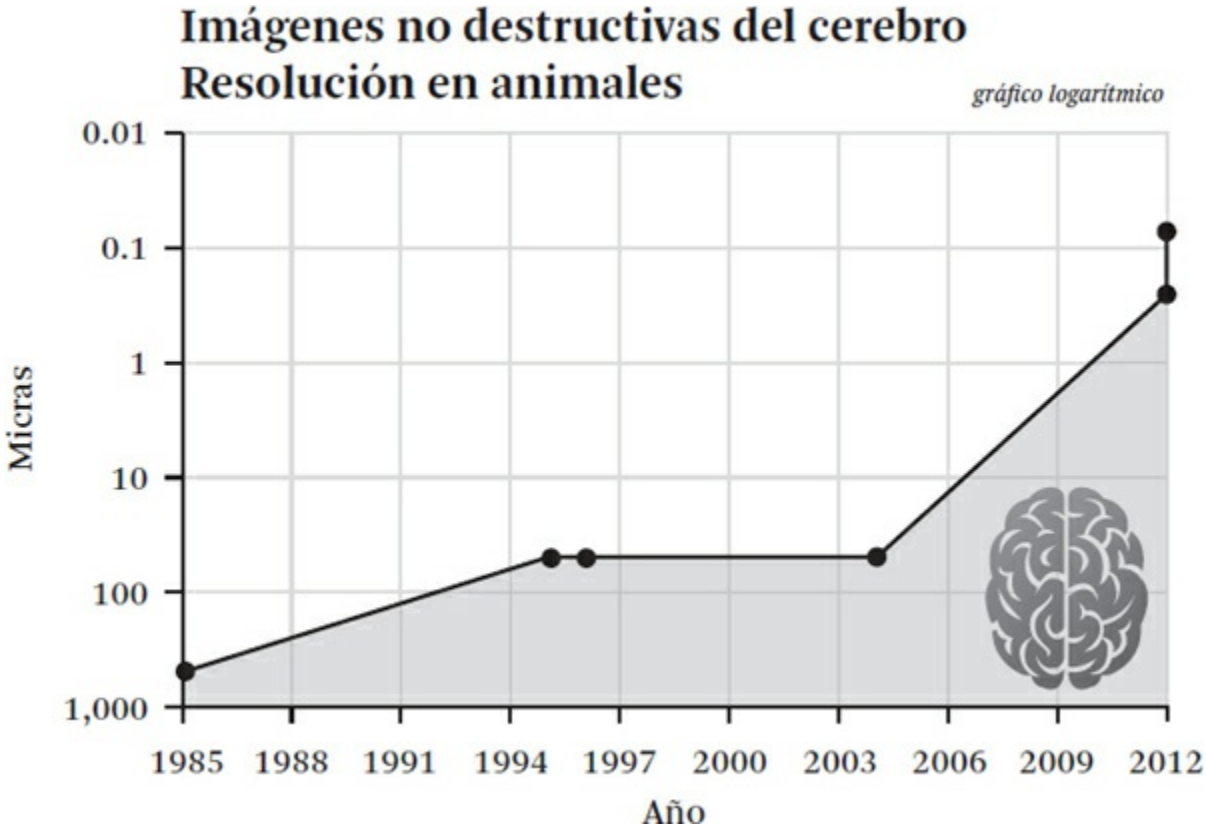

**Herramientas para tomar imágenes del cerebro (20).**




**Resolución espacial IRM en micras (21).**




**Resolución espacial de las técnicas destructivas por imágenes (22).**



**Resolución espacial de las técnicas no destructivas por imágenes en animales (23).**

Las tecnologías de inteligencia artificial tales como los sistemas para comprender el lenguaje natural no están necesariamente diseñadas para emular los principios teóricos del funcionamiento del cerebro, sino que están diseñadas para garantizar una máxima efectividad. Teniendo esto en cuenta, es de resaltar que las técnicas que han triunfado son coherentes con los principios que he subrayado en este libro: autoorganización, reconocedores jerárquicos de patrones autoasociativos invariables dotados de redundancia y predicciones en sentido ascendente y descendente. Estos sistemas también están desarrollándose exponencialmente, tal y como Watson ha demostrado.

Un objetivo primordial de la comprensión del cerebro es la mejora de las herramientas de las que constan las técnicas destinadas a crear sistemas inteligentes. Aunque es posible que muchos investigadores en el campo de IA no lo reconozcan del todo, dichas herramientas ya están profundamente influenciadas por nuestro conocimiento sobre los principios operativos del cerebro. Además, comprender el cerebro también nos ayuda a revertir disfuncionalidades cerebrales de varios tipos. Por supuesto, el proyecto de aplicar la ingeniería inversa al cerebro también persigue otro objetivo fundamental: comprender quienes somos.

* *Law of accelerated returns.*

CAPÍTULO ONCE

# Objeciones

> Si una máquina demuestra ser indistinguible de un humano, deberíamos dispensarle el mismo respeto que a un humano, deberíamos aceptar que posee una mente.
>
> —STEVAN HARNAD

El principal foco de objeciones a mi tesis de la ley de los rendimientos acelerados y sus aplicaciones para la amplificación de la inteligencia humana se encuentra en la naturaleza lineal de la intuición humana. Tal y como describí anteriormente, todos los varios cientos de millones de reconocedores de patrones del neocórtex procesan información secuencialmente. Una de las consecuencias de esta forma de organización es que sobre el futuro poseemos expectativas lineales. Así, los críticos aplican su intuición lineal a fenómenos de la información que fundamentalmente son exponenciales.

A este tipo de objeciones las llamo "críticas de incredulidad", ya que las proyecciones exponenciales parecen increíbles desde nuestra predilección lineal. Estas objeciones se presentan bajo varias formas. Paul Allen, cofundador de Microsoft nacido en 1953, y su colega Mark Greaves expusieron recientemente varias de estas objeciones en un ensayo titulado "The Singularity Isn't Near"* publicado en la revista *Technology Review* (1). Aunque mi respuesta en estas páginas va dirigida a las críticas de Allen en particular, sus críticas representan una variedad típica de objeciones a los argumentos que sostengo, sobre todo en lo que respecta al cerebro. Aunque el título de su ensayo hace referencia a *La Singularidad está cerca*, Allen solo cita un ensayo que escribí en 2001 ("La ley de los rendimientos acelerados"). Además, su artículo ni reconoce ni responde a los argumentos que esgrimo en el libro. Por desgracia, esto es algo habitual en las críticas que leo sobre mi trabajo.

Cuando en 1999 se publicó *The Age of Spiritual Machines*, libro sobre el que posteriormente profundicé en mi ensayo de 2001, se alzaron varias voces críticas del tipo: *la ley de Moore tocará a su fin; es posible que la capacidad del hardware se expanda exponencialmente pero el software está atascado en el lodo; el cerebro es demasiado complicado; el cerebro posee capacidades que el software es intrínsecamente incapaz de replicar;* etc. Tanto es así, que una de las razones por

las que escribí “La Singularidad está cerca” fue para responder a estas críticas.

No puedo decir que Allen y críticos similares tengan que estar necesariamente convencidos de que lo que argumento en el libro sea verdad, pero por lo menos él y otros podrían haber respondido a lo que escribí realmente. Allen sostiene que “la ley de los rendimientos acelerados (LOAR) […] no es una ley física”. A eso tengo que decir que la mayoría de las leyes científicas no son leyes físicas, sino el resultado de las propiedades emergentes de un gran número de acontecimientos a un nivel más bajo. Un ejemplo clásico son las leyes de la termodinámica. Si observamos las fórmulas matemáticas que subyacen tras las leyes de la termodinámica, veremos que modelizan las partículas como si estas dieran un paseo aleatorio, de manera que por definición no podemos predecir dónde se encontrará una partícula en particular en el futuro. Sin embargo, las propiedades generales de un gas son predecibles con un alto grado de precisión si se usan las *leyes* de la termodinámica. La ley de los rendimientos acelerados sostiene lo siguiente: todo proyecto y avance tecnológico es impredecible; sin embargo, la trayectoria general, cuantificada a través de mediciones fundamentales en lo que respecta a la relación entre rendimiento y precio, así como en lo que respecta a la capacidad, describen un sendero extraordinariamente predecible.

Si la tecnología informática solo estuviera siendo desarrollada por un puñado de investigadores, esta sí que sería impredecible. Sin embargo, esta tecnología es el resultado de un sistema suficientemente dinámico basado en proyectos competitivos, lo cual significa que una medida básica de la relación que se establece entre rendimiento y precio, como por ejemplo los cálculos por segundo por dólar constante, describe un sendero exponencial muy sólido que data del censo de EE.UU. del año 1890, tal y como he señalado en el capítulo anterior. Aunque las bases teóricas de la LOAR aparecen exhaustivamente explicadas en *La Singularidad está cerca*, la mayor defensa a su favor proviene de las amplias evidencias empíricas que otros y yo mismo hemos hecho patentes.

Allen escribe que “estas ‘leyes’ se cumplen hasta que dejan de hacerlo”. En este caso confunde los paradigmas con la trayectoria seguida en las áreas fundamentales de las tecnologías de la información. Si por ejemplo tuviéramos que estudiar la tendencia en la creación de tubos de vacío cada vez más pequeños (el paradigma de la mejora computacional en la década de 1950), es cierto que esta continuó hasta que dejó de hacerlo. Sin embargo, al mismo tiempo que el fin de este paradigma en particular se hacía evidente, la presión investigadora crecía para dar lugar al siguiente paradigma. La tecnología basada en los transistores mantuvo la tendencia exponencial del crecimiento en lo que se refiere a la relación entre rendimiento y precio de la computación, y esto dio lugar al quinto paradigma (la ley de Moore) y a la cada vez mayor compresión de los dispositivos en el interior de los circuitos integrados. Periódicamente se han escuchado predicciones que dicen que la ley de

Moore tocará a su fin. La industria de los semiconductores, en su "International Technology Roadmap for Semiconductors", prevé la aparición de dispositivos de 7 nanómetros a principios de la década de 2020 (2). En dicho momento, dispositivos fundamentales tendrán una anchura de 35 átomos de carbono y será difícil continuar reduciéndolos más. Sin embargo, Intel y otros productores de chips ya están dando los primeros pasos hacia el sexto paradigma, la computación en tres dimensiones, para continuar la mejora exponencial en la relación precio/rendimiento. Intel prevé que los chips tridimensionales se generalizarán en 10 años (los transistores tridimensionales y los chips de memoria en 3D ya han sido comercializados). El sexto paradigma hará que la LOAR continúe cumpliéndose en lo que respecta a la relación entre rendimiento y precio de los ordenadores hasta que más tarde durante este siglo 1 000 dólares en computación sean billones de veces más poderosos que el cerebro humano (3). (Aparentemente, Allen y yo estamos de acuerdo en cuanto al nivel de computación necesario para simular funcionalmente el cerebro humano) (4).

Después Allen continúa con el típico argumento de que el software no está progresando de la misma forma exponencial que el hardware. En *La Singularidad está cerca* me ocupo en detalle de esta cuestión y cito diferentes métodos para medir la complejidad y la capacidad del software que muestran un crecimiento exponencial similar al del hardware (5). Un estudio reciente ("Raport to the President and Congress, Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology", realizado por el President's Council of Advisors on Science and Technology) sostiene lo siguiente:

> Todavía más reseñable (y todavía menos comprendido) es el hecho de que en muchas áreas, *las mejoras en el rendimiento debido a mejoras en los algoritmos han superado ampliamente incluso a las drásticas mejoras en el rendimiento debidas al aumento en la velocidad de los procesadores* . Los algoritmos que utilizamos hoy en el reconocimiento del habla, en la traducción del lenguaje natural, en el ajedrez y en la planificación logística han evolucionado asombrosamente durante la última década […]. He aquí tan solo un ejemplo ofrecido por el Profesor Martin Grötschel del Konrad-Zuse-Zentrum für Informationstechnik Berlin. Grötschel, experto en el campo de la optimización, señala que un modelo para referenciar la planificación de la producción habría tardado en resolverse 82 años en 1988 mediante los ordenadores y los algoritmos de programación lineal de aquellos días. 15 años después, en 2003, el mismo modelo podía resolverse en apenas un minuto, lo que representa una mejora de más o menos 43 millones de veces. De esta mejora, más o menos 1000 veces se deberían al aumento en la velocidad de procesamiento, ¡mientras que más o menos 43 000 veces se deberían a mejoras en los algoritmos! Grötschel también cita una mejora algorítmica de más o menos 30 000 veces en la programación integrada mixta* entre 1991 y 2008. El diseño y análisis de algoritmos, así como el estudio de la inherente complejidad computacional de los problemas, son subcampos fundamentales de la ciencia de la computación".

Téngase en cuenta que la programación que según Grötschel se ha beneficiado de una mejora en el rendimiento de 43 millones de veces es la técnica matemática utilizada para optimizar la asignación de recursos a sistemas jerárquicos de memoria tales como los HHMM que he expuesto anteriormente. En *La Singularidad está cerca* cito muchos otros ejemplos similares a este (6).

En lo que respecta a la inteligencia artificial, Allen se apresura a descalificar a Watson, el invento de IBM. Su opinión es compartida por muchos críticos. Muchos de estos detractores no saben nada sobre Watson, aparte de que se trata de software ejecutado en un ordenador (pese al hecho de ser un ordenador cuyo núcleo consta de 720 procesadores que funcionan en paralelo). Allen escribe que sistemas como Watson "siguen siendo frágiles, los límites de su rendimiento vienen determinados por sus presuposiciones internas y por los algoritmos que los definen, no pueden hacer generalizaciones y frecuentemente dan respuestas sin sentido en áreas que no son específicamente las suyas".

Primero, observaciones similares las podemos hacer sobre los humanos. También señalaría que "las áreas específicas" de Watson incluyen la Wikipedia *entera* además de muchas otras bases de conocimiento, lo cual difícilmente puede considerarse una especialidad reducida. Watson utiliza un amplio espectro de conocimientos humanos y es capaz de manejar sutiles formas del lenguaje, incluyendo juegos de palabras, símiles y metáforas en literalmente todos los campos de interés para los humanos. No es perfecto, pero tampoco lo son los humanos, y resultó lo suficientemente bueno como para ganar en el *Jeopardy!* a los mejores jugadores humanos.

Allen sostiene que Watson fue configurado por científicos, quienes construyeron todos los enlaces a áreas específicas de conocimientos reducidos. Esto, simplemente, no es verdad. Aunque unas pocas áreas de los datos de Watson fueron programadas directamente, Watson adquirió la inmensa mayoría de su conocimiento por su cuenta leyendo documentos en lenguaje natural tales como Wikipedia. Esto representa su punto fuerte más importante, al igual que su capacidad para comprender el intrincado lenguaje de las preguntas de *Jeopardy!* (respuestas para las que hay que formular una pregunta).

Tal y como he señalado anteriormente, gran parte de las críticas hacia Watson dicen que funciona mediante probabilidades estadísticas y que no posee una comprensión "verdadera". Muchos lectores interpretan esto como si significara que Watson se limita a reunir estadísticas sobre secuencias de palabras. De hecho, en el caso de Watson el término "información estadística" se refiere a la distribución de coeficientes y conexiones simbólicas de los métodos autoorganizativos, como por ejemplo los modelos ocultos jerárquicos de Márkov. Se podrían despreciar de la misma manera las distribuciones en las concentraciones de neurotransmisores y en los redundantes patrones de conexiones del córtex humano, que también conforman "información estadística". De hecho, nosotros resolvemos las ambigüedades prácticamente igual a como lo hace Watson: considerando la probabilidad de las diferentes interpretaciones de una frase.

Allen continúa diciendo: "cada estructura [del cerebro] ha sido precisamente

moldeada por millones de años de evolución para realizar una cosa en particular, independientemente de lo que esto sea. No es como un ordenador, cuyos miles de millones de transistores idénticos contenidos por matrices regulares de memoria son controlados por una CPU dotada de unos pocos elementos. En el cerebro cada estructura individual y cada circuito neuronal ha sido individualmente mejorado por la evolución y los factores medioambientales".

Esto significa que cada estructura y circuito neuronal del cerebro es único y por lo tanto su reproducción es imposible, ya que esto significaría reproducir el anteproyecto del cerebro, cosa que requeriría de cientos de billones de bytes de información. El plan estructural del cerebro (al igual que el del resto del cuerpo) está contenido en el genoma, y el cerebro por sí solo no puede contener más información referente al diseño que el propio genoma. Téngase en cuenta que la información epigenética (como por ejemplo los péptidos que controlan la expresión génica) no contribuyen de forma apreciable a la cantidad de información contenida en el genoma. La experiencia y el aprendizaje sí que contribuyen considerablemente a la cantidad de información contenida en el cerebro, pero lo mismo puede ser dicho de sistemas de inteligencia artificial como Watson. En *La Singularidad está cerca* demuestro que, después de una comprensión sin pérdidas y debido a la masiva redundancia del genoma, la cantidad de información concerniente al diseño contenida por el genoma es de más o menos 50 millones de bytes, apenas la mitad de la cual (es decir, unos 25 millones de bytes) pertenece al cerebro (7). No se trata de algo simple, pero representa un nivel de complejidad con el que podemos manejarnos y además representa menos complejidad que muchos sistemas de software del mundo de hoy en día. Y no solo eso, gran parte de los 25 millones de bytes de información genética concerniente al cerebro se corresponden con las necesidades biológicas de las neuronas, no con sus algoritmos para el procesamiento de información.

¿Cómo llegar a los entre 100 y 1000 billones de conexiones cerebrales partiendo solamente de las decenas de millones de bytes correspondientes a la información concerniente al diseño? Obviamente, la respuesta es: a través de una redundancia masiva. Dharmendra Modha, gerente de *Cognitive Computing* para *IBM Research*, escribe que "los neuroanatomistas no se han encontrado con una red irremediablemente entrecruzada y arbitrariamente conectada que sea completamente idiosincrática con respecto al cerebro de cada individuo. En lugar de eso se han encontrado una gran cantidad de estructuras que se repiten en el interior del cerebro y un alto grado de homogeneidad entre especies […]. La increíble capacidad de reconfiguración natural [del cerebro] nos da esperanzas de que los algoritmos nucleares de la neurocomputación sean independientes de las específicas modalidades sensoriales o motoras responsables de gran parte de la variación observada en las áreas de la estructura cortical. Asimismo, esperamos que esto se

traduzca en una mejora del circuito canónico. Ciertamente, es este circuito canónico a lo que desearíamos poder aplicar la ingeniería inversa" (8).

Allen sostiene la existencia de un inherente "freno de complejidad que necesariamente limitaría nuestro progreso a la hora de comprender el cerebro humano y replicar sus capacidades" y para decir esto se basa en el hecho de que cada una de las entre 100 y 1000 billones de conexiones del cerebro humano existe gracias a un diseño que explícitamente así lo determina. Su "freno de complejidad" confunde el bosque con los árboles. Si lo que se quiere es comprender, modelizar, simular y recrear un páncreas, no es necesario recrear o simular cada orgánulo en cada isleta pancreática. En lugar de eso, basta con comprender una sola isleta y luego hacer una abstracción de su funcionamiento básico en lo que se refiere al control de la insulina para posteriormente extender este conocimiento al conjunto de este tipo de células. En lo que se refiere a las isletas de células, este algoritmo es bien conocido. Ya existen páncreas artificiales que utilizan este modelo funcional y que están siendo testados. Aunque ciertamente en el cerebro existe mucha más complejidad y variabilidad que en las extremadamente repetitivas isletas de células pancreáticas, también es cierto que en el cerebro se da una enorme cantidad de repetición de funciones, tal y como he descrito profusamente en este libro.

Las críticas como las de Allen también expresan lo que yo llamo el "pesimismo del científico". Los investigadores que trabajan en la siguiente generación de una tecnología o en la modelización de un área científica están condenados a luchar contra un conjunto de desafíos inmediatos. Así, si alguien les describe el aspecto que dicha tecnología adoptará tras diez generaciones, sus ojos, de repente, se vuelven vidriosos. Uno de los pioneros de los circuitos integrados me recordó recientemente la lucha por pasar de tamaños de 10 micras (10 000 nanómetro) a tamaños 5 micras (5 000 nanómetros) hace más de 30 años. Los científicos estaban moderadamente confiados en poder conseguir su objetivo, pero cuando otras personas predecían que algún día tendríamos circuitos con componentes cuyo tamaño estaría por debajo de 1 micra (1000 nanómetros), la mayoría de ellos, tan concentrados como estaban en lograr su objetivo, pensaba que una cosa así era demasiado aventurada como para ser verosímil. Por eso hacían objeciones con respecto a la fragilidad de los circuitos a ese nivel de precisión, a los efectos térmicos, etc. A día de hoy, Intel está empezando a usar chips con puertas cuya longitud es de 22 nanómetros.

Este mismo tipo de pesimismo lo presenciamos con respecto al Proyecto del Genoma Humano. A mitad de camino de los 15 años que duró el proyecto, solo el 1% del genoma había sido reunido y los críticos hablaban de posibles límites esenciales en cuanto a la velocidad a la que podría ser secuenciado sin destruir sus delicadas estructuras genéticas. Sin embargo, gracias al crecimiento exponencial tanto en la capacidad como en la relación precio/rendimiento, el proyecto concluyó 7

años antes de lo esperado. El proyecto de aplicar la ingeniería inversa al cerebro humano está haciendo progresos parecidos. Por ejemplo, solo recientemente hemos cruzado el umbral que nos permite ver la formación y la activación de conexiones interneuronales individuales utilizando técnicas de escaneo no invasivas. Gran parte de las evidencias que presento en en este libro se basan en dichos avances, que se han producido en en el pasado más inmediato.

Allen describe mi propuesta de aplicar la ingeniería inversa al cerebro humano como un mero escaneo del cerebro que nos permita comprender su estructura más íntima para después simular un cerebro entero "de abajo a arriba" sin necesidad de comprender sus métodos de procesamiento de información. Sin embargo, esto no es lo que yo propongo. Sí que necesitamos comprender detalladamente cómo funcionan los tipos individuales de neuronas para luego recopilar información sobre cómo se conectan los módulos funcionales. Los métodos funcionales derivados de este tipo de análisis pueden entonces guiarnos en el desarrollo de sistemas inteligentes. Básicamente, lo que estamos buscando son métodos de inspiración biológica que puedan acelerar nuestra labor en el campo de la inteligencia artificial (por cierto, campo en el que la mayor parte de los avances se han hecho sin tener un gran conocimiento sobre la manera en la que el cerebro realiza funciones parecidas). Así, gracias a mi trabajo en el reconocimiento del habla, sé que nuestro trabajo sufrió una gran aceleración cuando empezamos a conocer la manera en la que el cerebro trata y transforma la información auditiva.

La manera en la que las estructuras enormemente redundantes del cerebro se diferencian se basa en el aprendizaje y la experiencia. De hecho, el actual estado de cosas en el campo de la IA permite que los sistemas también aprendan a partir de sus propias experiencias. Los coches autoconducidos de Google aprenden a partir de sus propias experiencias de conducción, así como de datos procedentes de coches de Google conducidos por conductores humanos. Por su parte, Watson aprendió la mayor parte de sus conocimientos a partir de lecturas hechas por su cuenta. Es interesante resaltar que los métodos empleados a día de hoy por la IA han evolucionado hasta ser matemáticamente muy similares a los mecanismos del neocórtex.

Otra objeción con respecto a la viabilidad de la "IA fuerte" (inteligencia artificial de nivel humano y superior) que suele ser esgrimida es que el cerebro humano realiza un gran uso de la computación analógica, mientras que los métodos digitales son intrínsecamente incapaces de replicar las gradaciones de valor que las representaciones analógicas pueden abarcar. Es cierto que un bit está o bien encendido o bien apagado; sin embargo, palabras de múltiples bits representan fácilmente gradaciones múltiples y además lo hacen con el nivel de precisión que se desee. Por supuesto, esto lo hacen los ordenadores digitales todo el tiempo. Así, la precisión de la información analógica del cerebro (la fuerza sináptica, por ejemplo)

es más o menos la representada por un nivel de entre los 256 niveles posibles que pueden ser representados mediante 8 dígitos.

En el capítulo 9 hice referencia a la objeción de Roger Penrose y de Stuart Hameroff concerniente a los microtúbulos y la computación cuántica. Recuérdese que sostenían que las estructuras de microtúbulos en las neuronas realizan computación cuántica y que como no es posible conseguir tal cosa en los ordenadores, los cerebros humanos son básicamente diferentes a los ordenadores y presumiblemente mejores. Tal y como expuse anteriormente, no hay evidencias de que microtúbulos neuronales realicen computación cuántica. De hecho, los humanos son bastante malos a la hora de resolver los tipos de problemas en los que un ordenador cuántico sería sobresaliente, como por ejemplo la factorización de grandes números. Asimismo, si algo de esto resultara ser verdad, nada impediría usar computación cuántica en nuestros ordenadores.

John Searle se hizo famoso por presentar un experimento mental al que llama "la habitación china", una cuestión que discuto en detalle en *La Singularidad está cerca* (9). En pocas palabras, tiene que ver con un hombre que recibe preguntas escritas en chino y luego las responde. Para ello utiliza un complicado manual. Searle sostiene que el hombre no posee un verdadero conocimiento del chino y que no es "consciente" del lenguaje, ya que no comprende ni las cuestiones ni las respuestas pese a su aparente capacidad para contestar preguntas en chino. Searle compara esto con un ordenador y concluye que un ordenador que pudiera contestar preguntas en chino (lo que viene a significar que pasaría un test de Turing en chino) no poseería, igual que el hombre de la habitación china, ni una comprensión real del lenguaje, ni consciencia de lo que está haciendo.

En el argumento de Searle se dan un par de juegos de manos filosóficos. Por un lado, el hombre de su experimento mental solo puede ser comparado con la unidad central de procesamiento (CPU) de un ordenador. Se podría argumentar que una CPU no tiene una comprensión verdadera de lo que hace, pero la CPU solo es una parte de la estructura. En la habitación china de Searle es el hombre *con* el manual el que conforma la totalidad del sistema. Dicho sistema sí que comprende el chino, de lo contrario no podría contestar de forma convincente las preguntas en chino (cosa que violaría las premisas que Searle establece para este experimento mental).

Lo atractivo del argumento de Searle nace del hecho de que a día de hoy es difícil deducir la existencia de entendimiento y consciencia reales a partir de un programa de ordenador. Sin embargo, el problema de este argumento es que podemos aplicar la misma forma de razonar al propio cerebro humano. Cada reconocedor de patrones neocortical (de hecho, cada neurona y cada componente neuronal) se rige por un algoritmo, ya que después de todo se trata de mecanismos moleculares que se rigen por las leyes naturales. Si llegamos a la conclusión de que

seguir un algoritmo es incompatible con un entendimiento y consciencia verdaderos, entonces a su vez tendríamos que concluir que el cerebro humano tampoco hace gala de estas cualidades. Se puede tomar el razonamiento de la habitación china de John Searle y simplemente sustituir "la manipulación de las conexiones interneuronales y fuerzas sinápticas" por su propias palabras cuando hace referencia a "la manipulación de símbolos". De esta manera se conseguiría construir un argumento convincente para defender que el cerebro humano no puede poseer una verdadera comprensión de nada.

Otra línea de su razonamiento proviene de la naturaleza de la naturaleza, que se ha convertido en algo sagrado para muchos analistas. Por ejemplo, el biólogo neozelandés Michael Denton (nacido en 1943) percibe una profunda diferencia entre los principios del diseño de las máquinas y los de la biología. Denton escribe que las entidades naturales son "autoorganizativas,[…] autoreferenciales, […] autoreplicantes, […] recíprocas, […] autoformativas y […] holísticas" (10). Sostiene además que dichas formas biológicas solo pueden ser creadas por medio de procesos biológicos y que por tanto estas formas son realidades de la existencia "inmutables, […] impenetrables y […] fundamentales", y que por lo tanto constituyen una categoría filosófica fundamentalmente diferente a la de las máquinas.

Tal y como hemos visto, la realidad es que las máquinas pueden ser diseñadas mediante estos mismos principios. El aprender los paradigmas de diseño específicos pertenecientes a la entidad más inteligente de la naturaleza (el cerebro humano) es precisamente el objetivo del proyecto para aplicar la ingeniería inversa al cerebro. Tampoco es cierto que los sistemas biológicos sean completamente "holísticos", tal y como Denton los describe, ni que por el contrario las máquinas tengan que ser completamente modulares. Hemos identificado con claridad jerarquías en las unidades de funcionamiento de los sistemas naturales, especialmente en el cerebro. Además, los sistemas de IA hacen uso de métodos similares.

Me da la impresión de que muchos críticos no se darán por satisfechos hasta que los ordenadores no pasen el test de Turing de forma habitual. Sin embargo, ni siquiera ese umbral será una marca clara. No hay duda de que se producirá controversia sobre la validez de los test de Turing que se realicen. De hecho, es probable que yo mismo me encuentre entre los críticos que ataquen a aquellos que pretendan hacernos creer que han conseguido pasar el test por primera vez. En el momento en el que las discusiones sobre la validez del test de Turing pasado por un ordenador se serenen, hará mucho tiempo que los ordenadores hayan sobrepasado la inteligencia de humanos no mejorados.

Aquí quiero hacer énfasis en las palabras "no mejorados", ya que el mejoramiento es precisamente la razón por la cual estamos creando estos "niños mentales", para citar a Hans Moravec (11). La combinación del reconocimiento de patrones a nivel humano y la velocidad y precisión propias de los ordenadores dará

como resultado la aparición de capacidades muy poderosas. Sin embargo, no se trata de una invasión alienígena mediante máquinas inteligentes procedentes de Marte. Estamos creando estas herramientas para convertirnos a nosotros mismos en seres más inteligentes. Estoy convencido de que la mayoría de los analistas estarán de acuerdo conmigo en que esto es lo que hace única a la especie humana, el hecho de que construimos estas herramientas para aumentar nuestro propio alcance.

* Literalmente "La Singularidad no está cerca".
* "Mixed interger programming" en el original.

# Epílogo

> Caballeros, el panorama es bastante desalentador […]. El clima mundial está cambiando, los mamíferos están haciéndose con el control y nosotros tenemos un cerebro de más o menos el tamaño de una nuez.
>
> —CONVERSACIÓN ENTRE DINOSAURIOS, EN *THE FAR SIDE* DE GARY LARSON

La inteligencia puede definirse como la capacidad para resolver problemas mediante recursos limitados, entre los cuales un recurso fundamental es el tiempo. Así, la capacidad para resolver un problema más deprisa, como por ejemplo la forma de encontrar comida o la manera de evitar a un depredador, refleja una mayor capacidad intelectual. La evolución dio lugar a la inteligencia porque le era útil a la supervivencia (esto es un hecho que puede parecer obvio, pero no todo el mundo está de acuerdo). Tal y como nuestra especie ha hecho uso de ella, la inteligencia nos ha permitido no solo dominar el planeta, sino también mejorar continuamente la calidad de nuestra vida. Este último hecho tampoco le resulta evidente a todo el mundo, ya que a día de hoy la creencia de que la vida está empeorando se está expandiendo. Por ejemplo, una encuesta de Gallup hecha pública el 4 de mayo de 2011 reflejaba que solo "el 44 % de los norteamericanos creen que la juventud de hoy disfrutará de una vida mejor que la de sus padres" (1).

Si observamos las tendencias de amplio espectro, la esperanza de vida humana no solo se ha cuadruplicado durante el último milenio (y se ha más que doblado durante los últimos dos siglos) (2), sino que el PIB per cápita medido en dólares corrientes ha pasado de cientos de dólares en el año 1800 a miles de dólares hoy en día. Además, las tendencias son todavía más pronunciadas en el mundo en desarrollo (3). Hace un siglo solo existía un puñado de democracias, mientras que a día de hoy la democracia se ha convertido en la norma. Para obtener una perspectiva histórica sobre lo mucho que hemos avanzado, recomiendo leer *Leviatán* (Thomas Hobbes, 1651), en el que se describe "la vida del hombre" como "solitaria, pobre, canallesca, embrutecida y corta". Desde una perspectiva actual, el reciente libro llamado *Abundance* (2012), escrito por el fundador de la X-Prize Foundation (y cofundador junto conmigo de la Singularity University) Peter Diamandis y por el escritor científico Steven Kotler, documenta las extraordinarias maneras en las que la vida ha mejorado en todas sus facetas hasta llegar al día de hoy. El reciente libro de Steven Pinker titulado *The Better Angels of Our Nature: Why Violence Has*

*Declined* (2011) documenta meticulosamente el constante aumento de las relaciones pacíficas entre personas y entre pueblos. La abogado, emprendedora y autora norteamericana Martine Rothblatt (nacida en 1954) documenta la constante mejora en cuestión de derechos civiles. Para ello pone como ejemplo el hecho de que en solo unas pocas décadas el matrimonio homosexual ha pasado de no ser reconocido legalmente en ningún sitio del mundo a ser legalmente aceptado en un número de jurisdicciones que crece rápidamente (4).

Una de las razones principales por las que la gente cree que la vida está empeorando es porque nuestra información sobre los problemas en el mundo ha venido aumentando progresivamente. Si a día de hoy se produce una batalla en cualquier parte del planeta, la experimentamos casi como si estuviéramos allí. Durante la Segunda Guerra Mundial, decenas de miles de personas podían morir en una batalla y el público en general solo podía verlo a través de un borroso noticiario cinematográfico semanas después de que hubiera ocurrido. Durante la Primera Guerra Mundial, una pequeña élite podía leer en el periódico (que no tenía fotografías) sobre la manera en que el conflicto progresaba. Durante el S. XIX prácticamente nadie tenía acceso a las noticias de forma oportuna.

El avance que hemos realizado como especie gracias a nuestra inteligencia se refleja en la evolución de nuestro conocimiento, lo cual incluye nuestra tecnología y nuestra cultura. Un mayor número de nuestras tecnologías se está convirtiendo en tecnologías de la información, por lo que continúan progresando intrínsecamente de forma exponencial. Gracias a dichas tecnologías somos capaces de hacer frente a los grandes desafíos de la humanidad, como por ejemplo el mantenimiento del medioambiente, el sustento (incluidas la energía, la comida y el agua) de una población que va en aumento, la superación de enfermedades, el gran aumento de la longevidad humana y la eliminación de la pobreza. Solo mediante la expansión de nosotros mismos a través de tecnología inteligente podemos afrontar a una escala adecuada la complejidad necesaria para hacer frente a estos retos.

Estas tecnologías no son la vanguardia de una invasión inteligente que competirá contra nosotros y que en último término acabará por arrinconarnos. Desde el momento en que agarramos un palo para llegar hasta una rama más alta hemos usado herramientas que nos permiten aumentar nuestro alcance, tanto físico como mental. El hecho de que hoy podamos sacar de nuestro bolsillo un dispositivo que nos da acceso a gran parte del conocimiento humano pulsando unas pocas teclas nos expande más allá de lo que la mayoría de analistas podían imaginar hace tan solo unas pocas décadas. El “teléfono móvil” (las comillas se deben que este aparato es mucho más que un teléfono) de mi bolsillo es un millón de veces más barato y miles de veces más poderoso que el ordenador que todos los estudiantes y profesores del MIT compartíamos cuando yo estudiaba la carrera. Esto significa un aumento de varios miles de millones de veces en la relación rendimiento/precio durante los

últimos cuarenta años y es una progresión que volveremos a presenciar durante los próximos 25 años, cuando lo que solíamos meter en un edificio y ahora metemos en nuestro bolsillo seamos capaces de meterlo en una célula sanguínea.

Así es como convergeremos con la tecnología inteligente que estamos creando. Nanobots inteligentes en nuestro torrente sanguíneo mantendrán nuestros cuerpos biológicos saludables a nivel celular y molecular. Se meterán en nuestro cerebro de forma no invasiva a través de los capilares e interactuarán con nuestras neuronas biológicas, lo que aumentará directamente nuestra inteligencia. Esto no es algo tan futurista como pueda parecer. Ya existen dispositivos del tamaño de una célula sanguínea que pueden curar la diabetes tipo I en animales o detectar y destruir células cancerígenas en el torrente sanguíneo. Basándonos en la ley de los rendimientos acelerados, dentro de tres décadas estas tecnologías serán mil millones de veces más poderosas de lo que lo son hoy.

Yo ya considero a los dispositivos que uso y a la nube de recursos informáticos a la que están virtualmente conectados como extensiones de mí mismo, y me siento incompleto si me separan de estos extensores cerebrales. Por eso la huelga de un día llevada a cabo por Google, Wikipedia y miles de otros sitios web contra SOPA (Stop Online Piracy Act) el 18 de enero de 2012 fue tan extraordinaria: me sentí como si parte de mi cerebro estuviera de huelga (aunque yo y otros encontramos la forma de acceder a estos recursos online). También fue una impresionante demostración de poder político por parte de estos sitios web, ya que el proyecto de ley (que parecía destinado a la ratificación) fue instantáneamente retirado. Pero lo más importante es que demostró lo mucho que ya hemos externalizado partes de nuestro pensamiento en la nube de computación. Esta ya forma parte de quién soy. Una vez que nos acostumbremos a tener inteligencia no biológica en el interior de nuestros cerebros, este acrecentamiento (y la nube a la que estará conectado) continuará aumentando su capacidad exponencialmente.

La inteligencia que crearemos a partir de la aplicación de la ingeniería inversa al cerebro tendrá acceso a su propio código fuente y será capaz de mejorarse a sí misma rápidamente en el contexto de un ciclo acelerado de diseño iterativo. Aunque, tal y como hemos visto, en el cerebro humano existe un considerable grado de plasticidad, el cerebro posee una arquitectura relativamente fija que no puede ser significativamente modificada, y además posee una capacidad limitada. No podemos aumentar sus 300 millones de reconocedores de patrones hasta convertirlos por ejemplo en 400 millones si no es a través de medios no biológicos. Una vez que podamos conseguirlo, no habrá ninguna razón para que nos detengamos en un nivel de capacidad determinado. Podremos continuar hasta alcanzar los mil millones de reconocedores de patrones, o el billón.

A partir de una mejora cuantitativa se produce un avance cualitativo. El avance

evolutivo más importante del homo sapiens fue cuantitativo, a saber, el desarrollo de una amplia frente en la que dar acomodo a una mayor cantidad de neocórtex. Una mayor capacidad neocortical le permitió a esta nueva especie crear y contemplar pensamientos a niveles conceptuales más elevados, lo cual dio lugar a la creación de los diferentes campos del arte y de la ciencia. A medida que añadamos mayor cantidad de neocórtex de forma no biológica, es de esperar que se den niveles de abstracción más altos.

El matemático británico Irvin J. Good, colega de Alan Turing, escribió en 1965 que "la primera máquina ultrainteligente es la última invención que el hombre tendrá que realizar". Good la definía como una máquina que podía sobrepasar las "actividades intelectuales de cualquier hombre sin importar lo inteligente que fuera dicho hombre" y llegó a la conclusión de que "como el diseño de máquinas es una de dichas actividades intelectuales, una máquina ultrainteligente podría diseñar máquinas todavía mejores, por lo que sin duda se acabaría produciendo una 'explosión de inteligencia'".

La última invención que la evolución biológica necesitaba producir (el neocórtex) está guiando inevitablemente a la humanidad hacia la última invención que necesita realizar: la creación de máquinas verdaderamente inteligentes y a que el diseño de una sea la inspiración de otra. La evolución biológica continúa, pero la evolución tecnológica avanza millones de veces más rápido. Según la ley de los rendimientos acelerados, al final de este siglo seremos capaces de llevar la computación hasta los máximos niveles posibles permitidos por las leyes de la física en el campo de la informática (5). A la materia y energía organizada de esta manera lo llamamos "computronium", que analizado libra a libra es enormemente más poderoso que el cerebro humano. No se tratará de mera computación, sino que estará imbuida de los algoritmos inteligentes que constituirán los conocimientos humano-máquina. Con el tiempo, convertiremos la mayor parte de la masa y de la energía apropiada para ello de nuestro pequeño rincón de la galaxia en computronium. Entonces, para hacer que la ley de los rendimientos acelerados siga cumpliéndose, tendremos que expandirnos por el resto de la galaxia y del universo.

Si es cierto que la velocidad de la luz es un límite inexorable, entonces la colonización del universo nos llevará mucho tiempo, ya que el sistema estelar más cercano a La Tierra está a cuatro años luz de distancia. Si existieran medios, aunque fueran muy ingeniosos, para sobrepasar este límite, nuestra inteligencia y tecnología serían lo suficientemente poderosas como para sacar partido de ellos. Esta es una de las razones por la que recientemente se ha dado tanta importancia al hecho de que los muones que cruzaron los 730 kilómetros que separan el acelerador CERN situado en la frontera francosuiza y el laboratorio Gran Sasso situado en la parte central de Italia dieran la impresión de desplazarse más rápido que la velocidad de luz. Esta observación en concreto parece ser una falsa alarma. Sin embargo, existen otras

posibilidades para salvar este límite. Ni siquiera tendríamos que superar la velocidad de la luz si pudiéramos encontrar atajos que nos llevaran hasta lugares aparentemente lejanos a través de dimensiones espaciales situadas más allá de las tres dimensiones a las que estamos acostumbrados. Averiguar si podemos superar o salvar el límite establecido por la velocidad de la luz será la cuestión estratégica fundamental a la que se enfrente la civilización humano-máquina de principios del siglo XXII.

Los cosmólogos discuten sobre si el mundo se acabará por culpa del fuego (un *big crunch* comparable al *big bang*) o por culpa del hielo* (la muerte de las estrellas a medida que se propagan en una expansión eterna). Sin embargo, esta discusión no toma en consideración el poder de la inteligencia, como si el surgimiento de esta solo hubiera sido un secundario espectáculo de entretenimiento en la gran mecánica celestial que gobierna el universo. ¿Cuánto tiempo nos llevará expandir nuestra inteligencia en su forma no biológica a través del universo? Si nos fuera posible trascender la velocidad de la luz (y admito que se trata de un gran "si" condicional) por ejemplo usando agujeros de gusano a través del espacio (lo cual es compatible con la física tal y como la entendemos a día de hoy), esto podría ser conseguido en unos poco siglos. De lo contrario, tardaremos mucho más. En cualquiera de los dos escenarios, nuestro sino es despertar al universo para luego decidir inteligentemente cuál es su destino imbuyéndole de inteligencia humana en su forma no biológica.

* En mi opinión, el autor hace aquí una alusión implícita al poema de Robert Frost "Fire and Ice" (Fuego y Hielo). "Algunos dicen que el mundo se acabará en fuego,/ algunos dicen que en hielo./ Por lo que sé del deseo/ mi bando es el de los que optan por el fuego./ Pero si tuviera que pensármelo dos veces,/ creo saber lo suficiente sobre el odio/ como para decir que el hielo para destruir/ también es estupendo/ y sería suficiente".

# Notas

## Introducción

1. He aquí una frase de “Cien años de soledad”, de Gabriel García Márquez:

> Aureliano Segundo no tuvo conciencia de la cantaleta hasta el día siguiente, después del desayuno, cuando se sintió aturdido por un abejorreo que era entonces más fluido y alto que el rumor de la lluvia, y era Fernanda que se paseaba por toda la casa doliéndose de que la hubieran educado como una reina para terminar de sirvienta en una casa de locos, con un marido holgazán, idólatra, libertino, que se acostaba boca arriba a esperar que le llovieran panes del cielo, mientras ella se destroncaba los riñones tratando de mantener a flote un hogar emparapetado con alfileres, donde había tanto que hacer, tanto que soportar y corregir desde que amanecía Dios hasta la hora de acostarse, que llegaba a la cama con los ojos llenos de polvo de vidrio y, sin embargo, nadie le había dicho nunca buenos días, Fernanda, qué tal noche pasaste, Fernanda, ni le habían preguntado aunque fuera por cortesía por qué estaba tan pálida ni por qué despertaba con esas ojeras de violeta, a pesar de que ella no esperaba, por supuesto, que aquello saliera del resto de una familia que al fin y al cabo la había tenido siempre como un estorbo, como el trapito de bajar la olla, como un monigote pintado en la pared, y que siempre andaban desbarrando contra ella por los rincones, llamándola santurrona, llamándola farisea, llamándola lagarta, y hasta Amaranta, que en paz descanse, había dicho de viva voz que ella era de las que confundían el recto con las témporas, bendito sea Dios, qué palabras, y ella había aguantado todo con resignación por las intenciones del Santo Padre, pero no había podido soportar más cuando el malvado de José Arcadio Segundo dijo que la perdición de la familia había sido abrirle las puertas a una cachaca, imagínese, una cachaca mandona, válgame Dios, una cachaca hija de la mala saliva, de la misma índole de los cachacos que mandó el gobierno a matar trabajadores, dígame usted, y se refería a nadie menos que a ella, la ahijada del duque de Alba, una dama con tanta alcurnia que le revolvía el hígado a las esposas de los presidentes, una fijodalga de sangre como ella que tenía derecho a firmar con once apellidos peninsulares, y que era el único mortal en ese pueblo de bastardos que no se sentía emberenjenado frente a dieciséis cubiertos, para que luego el adúltero do su marido dijera muerto de risa que tantas cucharas y tenedores, y tantos cuchillos y cucharitas no era cosa de cristianos, sino de ciempiés, y la única que podía determinar a ojos cerrados cuándo se servía el vino blanco, y de qué lado y en qué copa, y cuándo se servía el vino rojo, y de qué lado y en qué copa, y no como la montuna de Amaranta, que en paz descanse, que creía que el vino blanco se servía de día y el vino rojo do noche, y la única en todo el litoral que podía vanagloriarse de no haber hecho del cuerpo sino en bacinillas de oro, para que luego el coronel Aureliano Buendía, que en paz descanse, tuviera el atrevimiento do preguntar con su mala bilis de masón de dónde había merecido ese privilegio, si era que olla no cagaba mierda, sino astromelias, imagínense, con esas palabras, y para que Renata, su propia hija, que por indiscreción había visto sus aguas mayores en el dormitorio, contestara que de verdad la bacinilla era de mucho oro y de mucha heráldica, pero que lo que tenía dentro era pura mierda, mierda física, y peor todavía que las otras porque era mierda de cachaca, imagínese, su propia hija, de modo que nunca se había hecho ilusiones con el resto de la familia, pero de todos modos tenía derecho a esperar un poco de más consideración de parto do su esposo, puesto que bien o mal era su cónyuge de sacramento, su autor, su legítimo perjudicador, que se echó encima por voluntad libre y soberana la grave responsabilidad de sacarla del solar paterno, donde nunca se privé ni se dolió de nada, donde tejía palmas fúnebres por gusto de entretenimiento, puesto que su padrino había mandado una carta con su firma y el sello de su anillo impreso en el lacre, sólo para decir que las manos de su ahijada no estaban hechas para menesteres de este mundo,

como no fuera tocar el clavicordio y, sin embargo, el insensato de su marido la había sacado de su casa con todas las admoniciones y advertencias y la había llevado a aquella paila de infierno donde no se podía respirar de calor, y antes de que ella acabara de guardar sus dietas de Pentecostés ya se había ido con sus baúles trashumantes y su acordeón de perdulario a holgar en adulterio con una desdichada a quien bastaba con verle las nalgas, bueno, ya estaba dicho, a quien bastaba con verle menear las nalgas de potranca para adivinar que era una, que era una, todo lo contrario de ella, que era una dama en el palacio o en la pocilga, en la mesa o en la cama, una dama de nación, temerosa de Dios, obediente de sus leyes y sumisa a su designio, y con quien no podía hacer, por supuesto, las maromas y vagabundinas que hacía con la otra, que por supuesto se prestaba a todo, como las matronas francesas, y peor aún, pensándolo bien, porque éstas al menos tenían la honradez de poner un foco colorado en la puerta, semejantes porquerías, imagínese, ni más faltaba, con la hija única y bienamada de doña Renata Argote y don Fernando del Carpio, y sobre todo de éste, por supuesto, un santo varón, un cristiano de los grandes, Caballero de la Orden del Santo Sepulcro, de esos que reciben directamente de Dios el privilegio de conservarse intactos en la tumba, con la piel tersa como raso de novia y los ojos vivos y diáfanos como las esmeraldas.

2. Véase el gráfico del capítulo 10 "Crecimiento del banco de genes. Datos sobre la secuenciación del ADN".

3. Cheng Zhang y Jianpeng Ma, "Enhanced Sampling and Applications in Protein Folding in Explicit Solvent," *Journal of Chemical Physics* 132, no. 24 (2010): 244101. Véase también http://folding.stanford.edu/English/About sobre el proyecto casero Folding@, que ha reunido más de cinco millones de ordenadores por todo el mundo para simular el pliegue de proteínas.

4. Para una descripción más complete de este tema, véase la sección "[El impacto…] sobre el destino inteligente del cosmos: por qué es probable que estemos solos en el universo" en el capítulo 6 de "La Singularidad está cerca", Ray Kurzweil (New York: Viking, 2005; Berlin: Lola Books, 2012).

5. James D. Watson, *Discovering the Brain* (Washington, DC: National Academies Press, 1992).

6. Sebastian Seung, *Connectome: How the Brain's Wiring Makes Us Who We Are* (New York: Houghton Mifflin Harcourt, 2012).

7 . "Mandelbrot Zoom," http://www.youtube.com/watch?v=gEw8xpb1a-RA; "Fractal Zoom Mandelbrot Corner," http://www.youtube.com/watch?v=G_GBwuYuOOs.

# Capítulo uno: Experimentos mentales históricos

1. Charles Darwin, *The Origin of Species* (P. F. Collier & Son, 1909), 185/95–96.

2. Darwin, *On the Origin of Species*, 751 (206. 1. 1-6), Peckham's Variorum edition, edited by Morse Peckham, *The Origin of Species by Charles Darwin: A Variorum Text* (Philadelphia: University of Pennsylvania Press, 1959).

3. R. Dahm, "Discovering DNA: Friedrich Miescher and the Early Years of Nucleic Acid Research," *Human Genetics* 122, no. 6 (2008): 565–81, doi:10.1007/s00439-007-0433-0; PMID 17901982.

4. Valery N. Soyfer, "The Consequences of Political Dictatorship for Russian Science," *Nature Reviews Genetics* 2, no. 9 (2001): 723–29, doi:10.1038/35088598; PMID 11533721.

5. J. D. Watson and F. H. C. Crick, "A Structure for Deoxyribose Nucleic Acid," *Nature* 171 (1953): 737–38, http://www.nature.com/nature/dna50/watsoncrick.pdf and "Double Helix: 50 Years of DNA," *Nature* archive, http://www.nature.com/nature/dna50/archive.html.

6. Franklin murió en 1958 y el Premio Nobel por el descubrimiento del ADN fue otorgado en 1962. El hecho de que también ella hubiera recibido el premio si hubiera estado viva en 1962 es un tema controvertido.

7. Albert Einstein, “On the Electrodynamics of Moving Bodies” (1905). Este trabajo puso las bases de la teoría de la relatividad especial. Véase Robert Bruce Lindsay and Henry Margenau, *Foundations of Physics* (Woodbridge, CT: Ox Bow Press, 1981), 330.

8. “Crookes radiometer,” Wikipedia, http://en.wikipedia.org/wiki/Crookes_radiometer.

9. Téngase en cuenta que parte del momento de los fotones se transfiere a las moléculas del aire en el interior de la bombilla (ya que no se trata de un vacío perfecto) y luego es transferido desde las moléculas de aire calentadas hasta la veleta.

10. Albert Einstein, “Does the Inertia of a Body Depend Upon Its Energy Content?” (1905). En este trabajo Einstein introduce la famosa fórmula $E = mc^2$.

11. “Albert Einstein’s Letters to President Franklin Delano Roosevelt,” http://hypertextbook.com/eworld/einstein.shtml.

# Capítulo tres: Un modelo del neocórtex. La teoría de la mente basada en el reconocimiento de patrones

1. Algunos no mamíferos como los cuervos, loros y pulpos se dice que son capaces de realizar cierto nivel de razonamientos; sin embargo, se trata de algo muy limitado y no es suficiente como para crear herramientas que sigan su propio curso evolutivo. Es posible que estos animales hayan adaptado otras regiones cerebrales para alcanzar un pequeño número de niveles de pensamiento jerárquico, pero es necesario un neocórtex para realizar el relativamente ilimitado pensamiento jerárquico que los humanos pueden llevar a cabo.

2. V. B. Mountcastle, “An Organizing Principle for Cerebral Function: The Unit Model and the Distributed System” (1978), in Gerald M. Edelman and Vernon B. Mountcastle, *The Mindful Brain: Cortical Organization and the Group-Selective Theory of Higher Brain Function* (Cambridge, MA: MIT Press, 1982).

3. Herbert A. Simon, “The Organization of Complex Systems,” in Howard H. Pattee, ed., *Hierarchy Theory: The Challenge of Complex Systems* (New York: George Braziller, Inc., 1973), http://blog.santafe.edu/wp-content/uploads/2009/03/simon1973.pdf.

4. Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch, “The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?” *Science* 298 (November 2002): 1569–79, http://www.sciencemag.org/content/298/5598/1569.short.

5. El siguiente pasaje del libro *Tracender: nueve pasos para vivir bien para siempre* (*Transcend: Nine Steps to Living Well Forever*), de Ray Kurzweil y Terry Grossman (New York: Rodale, 2009), describe esta técnica de sueño lúcido con más detalle:

   He desarrollado un método para resolver problemas mientras duermo. Lo he perfeccionado por mí mismo durante varias décadas y así he aprendido las sutilezas que hacen que funcione mejor.

   Empiezo por encomendarme un problema en el momento de meterme en la cama. Se puede tratar de cualquier clase de problema, un problema matemático, algo relacionado con alguno de mis inventos, una estrategia empresarial o incluso un problema interpersonal.

   Durante unos minutes reflexiono durante el problema, pero sin intentar resolverlo, ya que eso mataría la resolución creativa que está por aparecer. Sin embargo, sí que intento pensar sobre el problema, ¿qué es lo que sé sobre él?, ¿qué forma tomará una posible solución? Luego me voy a la cama. Estos preparativos hacen que mi mente subconsciente trabaje sobre el problema.

   **Terry**: Sigmund Freud señaló que cuando soñamos muchos de los censores de nuestro cerebro se relajan,

de manera que podemos soñar cosas que son socialmente, culturalmente o incluso sexualmente consideradas tabús. Podemos soñar cosas raras sobre las que no nos permitimos pensar durante el día. Esta es una de las razones por las cuales los sueños son tan extraños.

**Ray**: También existen anteojos profesionales que impiden pensar creativamente. El origen de muchos de ellos son el entrenamiento profesional, los bloqueos mentales del tipo "no puedes resolver un procesamiento de señales de esa manera" o del tipo "se supone que el lenguaje no debe usar esas reglas". Estas presuposiciones mentales también se relajan durante el sueño, de manera que se soñará sobre nuevas maneras de resolver problemas sin estar sujeto a estas constricciones diurnas.

**Terry**: También hay otra parte de nuestro cerebro que no trabaja durante el sueño, nuestras capacidades racionales para evaluar si una idea es razonable. Esta es otra de las razones por las cuales pasan cosas raras o fantásticas durante el sueño. Cuando el elefante atraviesa la pared, no nos asombramos sobre cómo lo ha hecho. Simplemente nos decimos a nosotros mismos: "un elefante a atravesado la pared, no es para tanto". De hecho, si de despierto en mitad de la noche suelo darme cuenta de que he estado soñando de manera extraña y sesgada sobre el problema que me he encomendado.

**Ray**: El siguiente paso se produce por la mañana durante el periodo de duermevela, al que se le suele llamar *sueño lúcido*. En este estado todavía se conservan los sentimientos e imágenes de los sueños, pero contando ya con las facultades racionales. Por ejemplo, te das cuenta de que estás en la cama, de manera que se puede formular el pensamiento racional que nos dice que tenemos mucho que hacer y que hay que abandonar la cama. Sin embargo, eso sería un error. Cuando puedo, me quedo en la cama y continúo en este estado de duermevela, ya que es fundamental para este método de resolución creativa de problemas. Por cierto, si la alarma se enciende esto no funciona.

**Lector**: Esto parece ser lo mejor de ambos mundos.

**Ray**: Exactamente. Sigo teniendo acceso a los pensamientos del sueño sobre el problema que me asigné la noche anterior, pero estoy lo suficientemente consciente y en un estado suficientemente racional como para evaluar las nuevas ideas creativas que se me ocurrieron durante la noche. Así, puedo decidir cuáles tienen sentido. Después de quizá 20 minutos siempre se me ocurren agudas ideas para el problema.

De esta manera se me han ocurrido inventos sobre los que me he pasado el resto del día escribiendo la patente, he organizado los materiales de este libro más satisfactoriamente y he tenido ideas útiles para una gran variedad de problemas. Si tengo que tomar una decisión importante siempre recorro este proceso. Después es muy probable que confíe en lo correcto de mi decisión.

El secreto es dejar a la mente a su aire, no ser inquisitivo y no preocuparse sobre el resultado de este método. Es lo contrario de la disciplina mental. Hay que pensar sobre el problema, pero dejando que las ideas surjan durante el sueño. Por la mañana hay que volver a dejar a la mente a su aire y revisar las extrañas ideas generadas por los sueños. He descubierto que se trata de un método de incalculable valor para sacar partido de la creatividad natural de mis sueños.

**Lector**: Los adictos al trabajo que se encuentran entre nosotros ya podemos trabajar durante el sueño. No estoy seguro de que esto le agrade a mi esposa.

**Ray**: De hecho, puedes tomártelo como si los sueños hicieran el trabajo por ti.

# Capítulo cuatro: El neocórtex biológico

1. Steven Pinker, *How the Mind Works* (New York: Norton, 1997), 152–53.

2. D. O. Hebb, *The Organization of Behavior* (New York: John Wiley & Sons, 1949).

3. Henry Markram and Rodrigo Perrin, "Innate Neural Assemblies for Lego Memory," *Frontiers in Neural Circuits* 5, no. 6 (2011).

4. Comunicación por email con Henry Markram, 19 de febrero de 2012.

5. Van J. Wedeen et al., "The Geometric Structure of the Brain Fiber Pathways," *Science* 335, no. 6076 (March 30, 2012).

6. Tai Sing Lee, "Computations in the Early Visual Cortex," *Journal of Physiology—Paris* 97 (2003): 121–39.

7. Una lista de trabajos puede ser encontrada en http://cbcl.mit.edu/people/poggio/tpcv_short_pubs.pdf.

8. Daniel J. Felleman and David C. Van Essen, "Distributed Hierarchical Processing in the Primate Cerebral Cortex," *Cerebral Cortex* 1, no. 1 (January/February 1991): 1–47. Un detallado análisis de las matemáticas bayesianas pertenecientes a las comunicaciones de arriba a abajo y de abajo a arriba en el neocórtex se encuentra en Tai Sing Lee in "Hierarchical Bayesian Inference in the Visual Cortex," *Journal of the Optical Society of America* 20, no. 7 (July 2003): 1434–48.

9. Uri Hasson et al., "A Hierarchy of Temporal Receptive Windows in Human Cortex," *Journal of Neuroscience* 28, no. 10 (March 5, 2008): 2539–50.

10. Marina Bedny et al., "Language Processing in the Occipital Cortex of Congenitally Blind Adults," *Proceedings of the National Academy of Sciences* 108, no. 11 (March 15, 2011): 4429–34.

11. Daniel E. Feldman, "Synaptic Mechanisms for Plasticity in Neocortex," *Annual Review of Neuroscience* 32 (2009): 33–55.

12. Aaron C. Koralek et al., "Corticostriatal Plasticity Is Necessary for Learning Intentional Neuroprosthetic Skills," *Nature* 483 (March 15, 2012): 331–35.

13. Comunicación por email con Randal Koene, enero de 2012.

14. Min Fu, Xinzhu Yu, Ju Lu, and Yi Zuo, "Repetitive Motor Learning Induces Coordinated Formation of Clustered Dendritic Spines *in Vivo," Nature* 483 (March 1, 2012): 92–95.

15. Dario Bonanomi et al., "Ret Is a Multifunctional Coreceptor That Integrates Diffusible-and Contact-Axon Guidance Signals," *Cell* 148, no. 3 (February 2012): 568–82.

16. Véase la nota 7 del capítulo 11.

# Capítulo cinco: El cerebro antiguo

1. Vernon B. Mountcastle, "The View from Within: Pathways to the Study of Perception," *Johns Hopkins Medical Journal* 136 (1975): 109–31.

2. B. Roska and F. Werblin, "Vertical Interactions Across Ten Parallel, Stacked Representations in the Mammalian Retina," *Nature* 410, no. 6828 (March 29, 2001): 583–87; "Eye Strips Images of All but Bare Essentials Before Sending Visual Information to Brain, UC Berkeley Research Shows," University of California at Berkeley news release, March 28, 2001, www.berkeley.edu/news/media/releases/2001/03/28_wers1.html.

3. Lloyd Watts, "Reverse-Engineering the Human Auditory Pathway," in J. Liu et al., eds., *WCCI 2012* (Berlin: Springer-Verlag, 2012), 47–59. Lloyd Watts, "Real-Time, High-Resolution Simulation of the Auditory Pathway, with Application to Cell-Phone Noise Reduction," *ISCAS* (June 2, 2010): 3821–24. Para otros trabajos véase http://www.lloydwatts.com/publications.html.

4. Véase Sandra Blakeslee, “Humanity? Maybe It’s All in the Wiring,” *New York Times*, December 11, 2003, http://www.nytimes.com/2003/12/09/science/09BRAI.html.

5. T. E. J. Behrens et al., “Non-Invasive Mapping of Connections between Human Thalamus and Cortex Using Diffusion Imaging,” *Nature Neuroscience* 6, no. 7 (July 2003): 750–57.

6. Timothy J. Buschman et al., “Neural Substrates of Cognitive Capacity Limitations,” *Proceedings of the National Academy of Sciences* 108, no. 27 (July 5, 2011):11252–55, http://www.pnas.org/content/108/27/11252.long.

7. Theodore W. Berger et al., “A Cortical Neural Prosthesis for Restoring and Enhancing Memory,” *Journal of Neural Engineering* 8, no. 4 (August 2011).

8. Las funciones base son funciones no lineales que pueden ser combinadas linealmente si se juntan funciones base con diferentes pesos. Así pueden aproximarse a cualquier función no lineal. A. Pouget y L. H. Snyder, “Computational Approaches to Sensorimotor Transformations,” *Nature Neuroscience* 3, no. 11 Supplement (November 2000): 1192–98.

9. J. R. Bloedel, “Functional Heterogeneity with Structural Homogeneity: How Does the Cerebellum Operate?” *Behavioral and Brain Sciences* 15, no. 4 (1992):666–78.

10. S. Grossberg and R. W. Paine, “A Neural Model of Cortico-Cerebellar Interactions during Attentive Imitation and Predictive Learning of Sequential Handwriting Movements,” *Neural Networks* 13, no. 8–9 (October–November 2000):999–1046.

11. Javier F. Medina and Michael D. Mauk, “Computer Simulation of Cerebellar Information Processing,” *Nature Neuroscience* 3 (November 2000):1205–11.

12. James Olds, “Pleasure Centers in the Brain,” *Scientifi c American* (October 1956):105–16. Aryeh Routtenberg, “The Reward System of the Brain,” *Scientific American* 239 (November 1978): 154–64. K. C. Berridge and M. L. Kringelbach, “Affective Neuroscience of Pleasure: Reward in Humans and Other Animals,” *Psychopharmacology* 199 (2008): 457–80. Morten L. Kringelbach, *The Pleasure Center: Trust Your Animal Instincts* (New York: Oxford University Press, 2009). Michael R. Liebowitz, *The Chemistry of Love* (Boston: Little, Brown, 1983). W. L. Witters and P. Jones-Witters, *Human Sexuality: A Biological Perspective* (New York: Van Nostrand, 1980).

# Capítulo seis: Habilidades transcendentes

1. Michael Nielsen, *Reinventing Discovery: The New Era of Networked Science* (Princeton, NJ: Princeton University Press, 2012), 1–3. T. Gowers and M. Nielsen, “Massively Collaborative Mathematics,” *Nature* 461, no. 7266 (2009): 879–81. “A Combinatorial Approach to Density Hales-Jewett,” *Gowers’s Weblog*, http://gowers.wordpress.com/2009/02/01/a-combinatorial-approach-to-density-halesjewett/. Michael Nielsen, “The Polymath Project: Scope of Participation,” March 20, 2009, http://michaelnielsen.org/blog/?p=584. Julie Rehmeyer, “SIAM: Massively Collaborative Mathematics,” Society for Industrial and Applied Mathematics, April 1, 2010, http://www.siam.org/news/news.php?id=1731.

2. P. Dayan and Q. J. M. Huys, “Serotonin, Inhibition, and Negative Mood,” *PLoS Computational Biology* 4, no. 1 (2008), http://compbiol.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pcbi.0040004

# Capítulo siete: El neocórtex digital de inspiración

# biológica

1. Gary Cziko, *Without Miracles: Universal Selection Theory and the Second Darwinian Revolution* (Cambridge, MA: MIT Press, 1955).

2. David Dalrymple ha sido mi ahijado desde que en 1999 cumplió ocho años. Sus antecedentes pueden ser encontrados aquí: http://esp.mit.edu/learn/teachers/davidad/bio.html, y aquí: http://www.brainsciences.org/Research-Team/mr-david-dalrymple.html.

3. Jonathan Fildes, "Artificial Brain '10 Years Away,' " BBC News, July 22, 2009, http://news.bbc.co.uk/2/hi/8164060.stm. Véase también el vídeo "Henry Markram on Simulating the Brain: The Next Decisive Years," http://www.kurzweilai.net/henry-markram-simulating-the-brain-next-decisive-years.

4. M. Mitchell Waldrop, "Computer Modelling: Brain in a Box," *Nature News*, February 22, 2012, http://www.nature.com/news/computer-modelling-brain-in-a-box-1.10066.

5. Jonah Lehrer, "Can a Thinking, Remembering, Decision-Making Biologically Accurate Brain Be Built from a Supercomputer?" *Seed*, http://seedmagazine.com/content/article/out_of_the_blue/.

6. Fildes, "Artifi cial Brain '10 Years Away.'"

7. Véase http://www.humanconnectomeproject.org/.

8. Anders Sandberg and Nick Bostrom, *Whole Brain Emulation: A Roadmap*, Technical Report # 2008–3 (2008), Future of Humanity Institute, Oxford University, www.fhi.ox.ac.uk/reports/2008-3.pdf.

9. He aquí el esquema básico del algoritmo de una red neuronal. Es posible introducir muchas variaciones sobre él y el diseñador del sistema tiene que proporcionar ciertos métodos y parámetros críticos que se detallan en las siguientes páginas.

   El crear una red neuronal a modo de solución para un problema conlleva los siguientes pasos:

   Definir el input.

   Definir la topología de la red neuronal (por ejemplo, las capas de neuronas y las conexiones entre neuronas).

   Entrenar a la red neuronal con ejemplos del problema.

   Ejecutar la red neuronal entrenada para resolver nuevos ejemplos del problema.

   Hacer pública la compañía de nuestra red neuronal.

   Excepto el último, estos pasos vienen detallados a continuación:

   **El input del problema**

   El input del problema de la red neuronal consiste en una serie de números. Este input puede ser:

   Una matriz bidimensional de números que representen los píxeles de una imagen en un sistema visual de reconocimiento de patrones; o

   (en un sistema de reconocimiento auditivo, por ejemplo del habla) una matriz bidimensional de números que representen un sonido y en la que la primera dimensión represente los parámetros del sonido (p. ej. Los componentes de frecuencia) y la segunda dimensión represente los diferentes puntos temporales; o

(en un sistema de reconocimiento de patrones arbitrario) una matriz $n$ dimensional de números que representen el patrón del input.

**Definición de la topología**

Para configurar la red neuronal, la arquitectura de cada neurona consiste en:

Múltiples inputs en los que cada input está "conectado" o bien al output de otra neurona o bien a uno de los números del input.

Por lo general, un solo output, que está conectado o bien al input de otra neurona (que suele encontrarse en una capa superior) o al output final.

*Configuración de la primera capa de neuronas*

Creación de $N_0$ neuronas en la primera capa. En cada una de estas neuronas, "se conectan" cada uno de los múltiples inputs de la neurona a "puntos" (p. ej. números) en el input del problema. Estas conexiones pueden ser determinadas de forma aleatoria o usando un algoritmo evolutivo (véase más abajo).

Asignación de una "fuerza sináptica" inicial para cada una de las conexiones creadas. Estos pesos pueden empezar siendo todos ellos los mismos, pueden ser asignados de forma aleatoria o pueden ser determinados de otra manera (véase más abajo).

*Configuración de las capas adicionales de neuronas*

Configuración de un total de $M$ capas de neuronas. En cada capa, configuración de las neuronas que se encuentren en ella. En el caso de $\text{capa}_i$:

Creación de $N_i$ neuronas en la $\text{capa}_i$. Para cada una de estas neuronas, "conectar" cada uno de los múltiples inputs de la neurona a los outputs de las neurona en la $\text{capa}_{i-1}$ (véanse las variaciones de más abajo).

Asignación de una "fuerza sináptica" inicial a cada una de las conexiones creadas. Estos pesos pueden empezar siendo los mismos, pueden ser asignados aleatoriamente o pueden ser determinados de otra manera (véase más abajo).

Los outputs de las neuronas en la $\text{capa}_M$ son los outputs de la red neuronal (véanse las variaciones más abajo).

*Pruebas de reconocimiento*

*Cómo funciona cada neurona*

Una vez que la neurona está configurada, esta hace lo siguiente en cada prueba de reconocimiento:

Cada input con peso en la neurona es computado mediante la multiplicación del output de la otra neurona (o input inicial) a la que el input de esta neurona está conectado mediante la fuerza sináptica de dicha conexión.

Se suman todos estos inputs con peso que le llegan a la neurona.

Si el resultado de la suma es superior al umbral de disparo de esta neurona, entonces se considera que esta neurona se dispara y que tiene un output de 1. Si no, su output es 0 (véanse las variaciones más abajo).

*Hágase lo siguiente en cada prueba de reconocimiento*

En cada capa, desde $\text{capa}_0$ hasta $\text{capa}_M$:

En cada neurona de la capa:

Súmense los pesos de sus inputs (cada peso = al output de la otra neurona [o input inicial] al que el input que llega a esta neurona está conectado multiplicado por la fuerza sináptica de dicha conexión).

Si la suma de los pesos de los inputs es mayor que el umbral de disparo de esta neurona, fíjese el output de esta neurona = 1, de otra manera fíjese en 0.

**Entrenamiento de las redes neuronales**

Ejecutar repetidas pruebas de reconocimiento en problemas de ejemplo.

Después de cada prueba, ajustar las fuerzas sinápticas de todas las conexiones interneuronales para mejorar el rendimiento de la red neuronal durante la prueba (véase la discusión de más abajo sobre cómo hacer esto).

Continuar con este entrenamiento hasta que el grado de precisión de la red neuronal deje de mejorar (p. ej. al alcanzar una asíntota).

*Decisiones fundamentales sobre el diseño*

En el sencillo esquema de más arriba, el diseñador de este algoritmo de red neuronal tiene que determinar inicialmente:

Lo que los números del input representan.

El número de capas de neuronas.

El número de neuronas de cada capa. (Cada capa tiene que tener necesariamente el mismo número de neuronas).

El número de inputs a cada neurona en cada capa. El número de inputs (p. ej. las conexiones interneuronales) también pueden variar de neurona a neurona y de capa a capa.

"El cableado" (p. ej. las conexiones). Para cada neurona en cada capa, esto consiste en una lista de otras neuronas, los outputs que constituyen los inputs de esta neurona. Esto es un área de diseño fundamental. Existen diferentes maneras de hacer esto:

1. Cablear la red neuronal aleatoriamente; o
2. Usar un algoritmo evolutivo (véase más abajo) para determiner un cableado óptimo; o
3. Usar el sistema que el diseñador considere como el mejor para determinar el cableado.

Las fuerzas sinápticas iniciales (p. ej. los pesos) de cada conexión. Existen diferentes maneras de hacer esto:

1. Fijar las fuerzas sinápticas en el mismo valor; o
2. Fijar las fuerzas sinápticas en valores diferentes y aleatorios; o
3. Usar un algoritmo evolutivo para determinar un conjunto óptimo de valores iniciales; o
4. Usar el sistema que el diseñador considere el más oportuno para determinar los valores iniciales.

El umbral de disparo de cada neurona.

Determinación del output. El output puede ser:

1. los outputs de la $capa_M$ de neuronas; o
2. el output de una sola neurona del output, cuyos inputs son los outputs de las neuronas en la $capa_M$; o
3. una función de (p. ej. la suma de) los outputs de las neuronas en la $capa_M$ o
4. otra función de los outputs neuronales de diferentes capas.

Determinar la manera en la que las fuerzas sinápticas de todas las conexiones se ajustan durante el entrenamiento de esta red neuronal. Esta es una decisión fundamental sobre el diseño y está sujeta a una gran cantidad de investigación y discusión. Existen diferentes maneras de llevar esto a cabo:

1. En cada prueba de reconocimiento, aumentar o disminuir las fuerzas sinápticas (por lo general poco) para que el output de la red neuronal se corresponda mejor con la respuesta correcta. Una manera de conseguir esto probar tanto un incremento como una disminución y ver cuál de las dos opciones es la mejor. Esto puede llevar tiempo, de manera que existen otros métodos para tomar decisiones locales sobre si aumentar o disminuir cada una de las fuerzas sinápticas.
2. Hay otros métodos estadísticos para modificar las fuerzas sinápticas después de cada prueba de reconocimiento, de manera que el rendimiento de la red neuronal en una prueba concreta se ajuste mejor a la solución correcta.

Téngase en cuenta que el entrenamiento de la red neuronal funcionará incluso si las respuestas a las pruebas de entrenamiento no son todas correctas. Esto permite utilizar material de entrenamiento perteneciente al mundo real y que contenga una inherente tasa de error. La cantidad de datos utilizados para el entrenamiento es un facto fundamental para el éxito de la red neuronal. Normalmente se necesita una gran cantidad para obtener resultados satisfactorios. Al igual que con estudiantes humanos, la cantidad de tiempo que una red neuronal se pasa aprendiendo sus lecciones es un factor fundamental en su rendimiento.

*Variaciones*

Se pueden hacer muchas variaciones sobre lo anterior. Por ejemplo:

Hay diferentes formas de determinar la topología. En concreto, el cableado interneuronal puede ser fijado o bien aleatoriamente o bien usando un algoritmo evolutivo.

Existen diferentes maneras de fijar las fuerzas sinápticas iniciales.

Los inputs de las neuronas en la $capa_i$ no tienen por qué proceder de los outputs de las neuronas de la $capa_{i-1}$. Los inputs de las neuronas de cada capa pueden proceder de cualquier capa inferior o superior.

Existen diferentes maneras de determinar el output final.

El método descrito más arriba resulta en un disparo de “todo o nada” (1 o 0) llamado no linealidad. Se pueden usar otras funciones no lineales. Por lo general, se usa una función que va del 0 al 1 de forma rápida pero más gradual. Los outputs también pueden numerarse con otros números que no sean el 0 y el 1.

Los diferentes métodos para ajustar las fuerzas sinápticas durante el entrenamiento representan decisiones fundamentales en cuanto al diseño.

El esquema de arriba describe una red neuronal “sincrónica” en la que cada prueba de reconocimiento se produce computando los outputs de cada capa, empezando por la $capa_0$ y terminando en la $capa_M$. En un sistema verdaderamente paralelo, en el que cada neurona opera independientemente de las otras, las neuronas pueden operar “asincrónicamente” (p. ej. independientemente). En la estrategia asincrónica cada neurona escanea constantemente sus inputs y se dispara cuando la suma de los pesos de sus inputs supera su umbral (o su función de output específica).

10. Robert Mannell, “Acoustic Representations of Speech,” 2008, http://clas.mq.edu.au/acoustics/frequency/acoustic_speech.html.

11. He aquí el esquema básico de un algoritmo genético (evolutivo). Se pueden realizar muchas variantes y el diseñador del sistema tiene que proporcionar ciertos parámetros y métodos fundamentales que se detallan más abajo.

**El algoritmo evolutivo**

Crear *N* "criaturas" solución. Cada una posee:

Un código genético: una secuencia de números que caracteriza una posible solución del problema. Los números pueden representar parámetros críticos, pasos de una solución, reglas, etc.

Para cada generación de la evolución, hágase lo siguiente:

Hágase lo siguiente para cada una de las *N* criaturas solución:

Aplicar al problema o al medio simulado la solución de esta criatura solución (tal y como viene representada por su código genético). Califique la solución.

Escoja las *L* criaturas solución con las mejores calificaciones para sobrevivir en la siguiente generación.

Elimine las (*N* – *L*) criaturas solución no supervivientes.

Crear (*N* – *L*) criaturas solución nuevas de entre las *L* criaturas solución supervivientes mediante:

1. Realización de copias de las *L* criaturas supervivientes Introducir pequeñas variantes aleatorias en cada copia; o:
2. Crear criaturas solución adicionales combinando partes del código genético (usando reproducción "sexual" o combinando partes de los cromosomas) a partir de la L criaturas supervivientes; o
3. Combinar (1) y (2).

Determinar si se sigue evolucionando o no:

Mejora = (calificación más alta de esta generación) – (calificación más alta de la generación anterior).

Si mejora < umbral de mejora, entonces hemos terminado.

La criatura solución con mejor calificación en la última generación de la evolución posee la mejor solución. Aplicar la solución definida por su código genético al problema.

*Decisiones fundamentales respecto al diseño*

En el sencillo esquema de más arriba, el diseñador tiene que determinar en un principio:

Parámetros fundamentales:

*N*

*L*

Umbral de mejora.

Lo que representan los números del código genético y cómo la solución es computada a partir del código genético.

Un método para determinar las N criaturas solución en la primera generación. En general, estas solo necesitan ser intentos de solución "razonables". Si estas soluciones de primera generación son demasiado imprecisas, el algoritmo evolutivo puede tener dificultades para llegar hasta la solución correcta. A menudo merece la pena crear las criaturas solución iniciales de manera que sean razonablemente diversas. Esto ayudará a prevenir que el proceso evolutivo se limite a encontrar una solución óptima "localmente".

Cómo se califican las soluciones.

Cómo se reproducen las criaturas solución supervivientes.

*Variaciones*

Se pueden hacer muchas variaciones sobre lo dicho más arriba. Por ejemplo:

No tiene por qué haber un número fijo de criaturas solución supervivientes ($L$) procedentes de cada generación. La(s) regla(s) de supervivencia puede(n) permitir un número variable de supervivientes.

No tiene por qué haber un número fijo de criaturas solución creadas en cada generación ($N - L$). Las reglas de procreación pueden ser independientes del tamaño de la población. La procreación puede estar relacionada con la supervivencia, permitiendo así que las criaturas solución sean las que más se procreen.

La decisión sobre si continuar o no la evolución puede variar. Se pueden tomar en consideración más criaturas solución, no solo las que tengan mejores calificaciones en la generación más reciente. También se puede tomar en consideración una tendencia que vaya más allá de las últimas dos generaciones.


12. Dileep George, "How the Brain Might Work: A Hierarchical and Temporal Model for Learning and Recognition" (PhD dissertation, Stanford University, June 2008).

13. A. M. Turing, "Computing Machinery and Intelligence," *Mind*, October 1950.

14. Hugh Loebner organiza cada año la competición "Loebner Prize". La medalla de plata de Loebner irá a parar al ordenador que pase el test de Turing original basado solamente en texto. La medalla de oro irá a parar al ordenador que pueda pasar una versión del test que incluya input de audio y video. En mi opinión, la inclusión del audio y del vídeo no hace que el test sea más difícil.

15. "Cognitive Assistant That Learns and Organizes," Artificial Intelligence Center, SRI International, http://www.ai.sri.com/project/CALO.

16. Dragon Go! Nuance Communications, Inc., http://www.nuance.com/products/dragon-go-in-action/index.htm.

17. "Overcoming Artificial Stupidity," *WolframAlpha Blog*, April 17, 2012, http://blog.wolframalpha.com/author/stephenwolfram/.


# Capítulo ocho: La mente como ordenador


1. Salomon Bochner, *A Biographical Memoir of John von Neumann* (Washington, DC: National Academy of Sciences, 1958).

2. A. M. Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem," *Proceedings of the London Mathematical Society* Series 2, vol. 42 (1936–37):230–65, http://www.comlab.ox.ac.uk/activities/ieg/e-library/sources/tp2-ie.pdf. A. M. Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem: A Correction," *Proceedings of the London Mathematical Society* 43 (1938): 544–46.

3. John von Neumann, "First Draft of a Report on the EDVAC," Moore School of Electrical Engineering, University of Pennsylvania, June 30, 1945. John von Neumann, "A Mathematical Theory of Communication," *Bell System Technical Journal*, July and October 1948.

4. Jeremy Bernstein, *The Analytical Engine: Computers—Past, Present, and Future*, rev. ed. (New York: William Morrow & Co., 1981).

5. "Japan's K Computer Tops 10 Petaflop/s to Stay Atop TOP500 List," *Top 500*, November 11, 2011, http://top500.org/lists/2011/11/press-release.

6. Carver Mead, *Analog VLSI and Neural Systems* (Reading, MA: Addison-Wesley, 1986).

7 . "IBM Unveils Cognitive Computing Chips," IBM news release, August 18, 2011, http://www-03.ibm.com/press/us/en/pressrelease/35251.wss.

8. "Japan's K Computer Tops 10 Petaflop/s to Stay Atop TOP500 List."

# Capítulo nueve: Experimentos mentales sobre la mente

1. John R. Searle, "I Married a Computer," in Jay W. Richards, ed., *Are We Spiritual Machines? Ray Kurzweil vs. the Critics of Strong AI* (Seattle: Discovery Institute, 2002).

2. Stuart Hameroff, *Ultimate Computing: Biomolecular Consciousness and Nanotechnology* (Amsterdam: Elsevier Science, 1987).

3. P. S. Sebel et al., "The Incidence of Awareness during Anesthesia: A Multicenter United States Study," *Anesthesia and Analgesia* 99 (2004): 833–39.

4. Stuart Sutherland, *The International Dictionary of Psychology* (New York: Macmillan, 1990).

5. David Cockburn, "Human Beings and Giant Squids," *Philosophy* 69, no. 268 (April 1994): 135–50.

6. Ivan Petrovich Pavlov, from a lecture given in 1913, published in *Lectures on Conditioned Reflexes: Twenty-Five Years of Objective Study of the Higher Nervous Activity [Behavior]of Animals* (London: Martin Lawrence, 1928), 222.

7. Roger W. Sperry, from James Arthur Lecture on the Evolution of the Human Brain, 1964, p. 2.

8. Henry Maudsley, "The Double Brain," *Mind* 14, no. 54 (1889): 161–87.

9. Susan Curtiss and Stella de Bode, "Language aft er Hemispherectomy," *Brain and Cognition* 43, nos. 1–3 (June–August 2000): 135–38.

10. E. P. Vining et al., "Why Would You Remove Half a Brain? The Outcome of 58 Children after Hemispherectomy— the Johns Hopkins Experience: 1968 to 1996," Pediatrics 100 (August 1997): 163–71. M. B. Pulsifer et al., "The Cognitive Outcome of Hemispherectomy in 71 Children," *Epilepsia* 45, no. 3 (March 2004):243–54.

11. S. McClelland III and R. E. Maxwell, "Hemispherectomy for Intractable Epilepsy in Adults: The First Reported Series," *Annals of Neurology* 61, no.4 (April 2007):372–76.

12. Lars Muckli, Marcus J. Naumerd, and Wolf Singer, "Bilateral Visual Field Maps in a Patient with Only One Hemisphere," *Proceedings of the National Academy of Sciences* 106, no. 31 (August 4, 2009), http://dx.doi.org/10.1073/pnas.0809688106.

13. Marvin Minsky, *The Society of Mind* (New York: Simon and Schuster, 1988).

14. F. Fay Evans-Martin, *The Nervous System* (New York: Chelsea House, 2005), http://www.scribd.com/doc/5012597/The-Nervous-System.

15. Benjamin Libet, *Mind Time: The Temporal Factor in Consciousness* (Cambridge, MA: Harvard University Press, 2005).

16. Daniel C. Dennett, *Freedom Evolves* (New York: Viking, 2003).

17. Michael S. Gazzaniga, *Who's in Charge? Free Will and the Science of the Brain* (New York: Ecco/ HarperCollins, 2011).

18. David Hume, *An Enquiry Concerning Human Understanding* (1765), 2nd ed., edited by Eric Steinberg (Indianapolis: Hackett, 1993).

19. Arthur Schopenhauer, *The Wisdom of Life*.

20. Arthur Schopenhauer, *On the Freedom of the Will* (1839).

21. From Raymond Smullyan, *5000 B. C. and Other Philosophical Fantasies* (New York: St. Martin's Press, 1983).

22. Para un profundo y entretenido examen de cuestiones similares relacionadas con la identidad y la consciencia, véase Martine Rothblatt, "The Terasem Mind Uploading Experiment," *International Journal of Machine Consciousness* 4, no. 1 (2012): 141–58. En este trabajo, Rothblatt examina la cuestión de la identidad con respecto al software que emule a una persona basándose en "una base de datos compuesta de entrevistas grabadas en video y de informaciones asociadas a la persona originaria". En este exprerimiento futuro en concreto, el software consigue emular a la persona en la que se basa.

23. "How Do You Persist When Your Molecules Don't?" *Science and Consciousness Review* 1, no. 1 (June 2004), http://www.sci-con.org/articles/20040601.html.

# Capítulo diez: La ley de los rendimientos acelerados aplicada al cerebro


1. "DNA Sequencing Costs," National Human Genome Research Institute, NIH, http://www.genome.gov/sequencingcosts/.

2. "Genetic Sequence Data Bank, Distribution Release Notes," December 15, 2009, National Center for Biotechnology Information, National Library of Medicine, ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt.

3. "DNA Sequencing— The History of DNA Sequencing," January 2, 2012, http://www.dnasequencing.org/history-of-dna.

4. "Cooper's Law," ArrayComm, http://www.arraycomm.com/technology/coopers-law.

5. "The Zettabyte Era," Cisco, http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/VNI_Hyperconnectivity_WP and "Number of Internet Hosts," Internet Systems Consortium, http://www.isc.org/solutions/survey/history.

6. TeleGeography © PriMetrica, Inc., 2012.

7. Dave Kristula, "The History of the Internet" (March 1997, update August 2001), http://www.davesite.com/webstation/net-history.shtml; Robert Zakon, "Hobbes' Internet Timeline v8. 0," http://www.zakon.org/robert/internet/timeline; Quest Communications, 8-K for 9/ 13/1998 EX-99. 1; *Converge! Network Digest*, December 5, 2002, http://www.convergedigest.com/Daily/daily.asp?vn=v9n229&fecha=December%2005,%202002; Jim Duff y, "AT& T Plans Backbone Upgrade to 40G," *Computerworld*, June 7, 2006, http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9001032; "40G: The Fastest Connection You Can Get?" InternetNews.com, November 2, 2007, http://www.internetnews.com/infra/article.php/3708936; "Verizon First Global Service Provider to Deploy 100G on U. S. Long-Haul Network," news release, Verizon, http://newscenter.verizon.com/press-releases/verizon/2011/verizon-first-global-service.html.

8. Facebook, "Key Facts," http://newsroom.fb.com/content/default.aspx?NewsAreaId=22.

9. http://www.kurzweilai.net/how-my-predictions-are-faring.

10. Cálculos por Segundo por $1,000

| *Año* | *Cálculos por Segundo por $1,000* | *Máquina* | *Logaritmo natural (calcs/sec/$k)* |
|---|---|---|---|
| 1900 | 5.82E−06 | Analytical Engine | −12.05404 |
| 1908 | 1.30E−04 | Hollerith Tabulator | −8.948746 |
| 1911 | 5.79E−05 | Monroe Calculator | −9.757311 |
| 1919 | 1.06E−03 | IBM Tabulator | −6.84572 |
| 1928 | 6.99E−04 | National Ellis 3000 | −7.265431 |
| 1939 | 8.55E−03 | Zuse | −4.762175 |
| 1940 | 1.43E−02 | Bell Calculator Model 1 | −4.246797 |
| 1941 | 4.63E−02 | Zuse 3 | −3.072613 |
| 1943 | 5.31E+00 | Colossus | 1.6692151 |
| 1946 | 7.98E−01 | ENIAC | −0.225521 |
| 1948 | 3.70E−01 | IBM SSEC | −0.994793 |
| 1949 | 1.84E+00 | BINAC | 0.6081338 |
| 1949 | 1.04E+00 | EDSAC | 0.0430595 |
| 1951 | 1.43E+00 | Univac I | 0.3576744 |
| 1953 | 6.10E+00 | Univac 1103 | 1.8089443 |
| 1953 | 1.19E+01 | IBM 701 | 2.4748563 |
| 1954 | 3.67E−01 | EDVAC | −1.002666 |
| 1955 | 1.65E+01 | Whirlwind | 2.8003255 |
| 1955 | 3.44E+00 | IBM 704 | 1.2348899 |
| 1958 | 3.26E−01 | Datamatic 1000 | −1.121779 |
| 1958 | 9.14E−01 | Univac II | −0.089487 |
| 1960 | 1.51E+00 | IBM 1620 | 0.4147552 |

| | | | |
|---|---|---|---|
| 1960 | 1.52E+02 | DEC PDP-1 | 5.0205856 |
| 1961 | 2.83E+02 | DEC PDP-4 | 5.6436786 |
| 1962 | 2.94E+01 | Univac III | 3.3820146 |
| 1964 | 1.59E+02 | CDC 6600 | 5.0663853 |
| 1965 | 4.83E+02 | IBM 1130 | 6.1791882 |
| 1965 | 1.79E+03 | DEC PDP-8 | 7.4910876 |
| 1966 | 4.97E+01 | IBM 360 Model 75 | 3.9064073 |
| 1968 | 2.14E+02 | DEC PDP-10 | 5.3641051 |
| 1973 | 7.29E+02 | Intellec-8 | 6.5911249 |
| 1973 | 3.40E+03 | Data General Nova | 8.1318248 |
| 1975 | 1.06E+04 | Altair 8800 | 9.2667207 |
| 1976 | 7.77E+02 | DEC PDP-11 Model 70 | 6.6554404 |
| 1977 | 3.72E+03 | Cray 1 | 8.2214789 |
| 1977 | 2.69E+04 | Apple II | 10.198766 |
| 1979 | 1.11E+03 | DEC VAX 11 Model 780 | 7.0157124 |
| 1980 | 5.62E+03 | Sun-1 | 8.6342649 |
| 1982 | 1.27E+05 | IBM PC | 11.748788 |
| 1982 | 1.27E+05 | Compaq Portable | 11.748788 |
| 1983 | 8.63E+04 | IBM AT-80286 | 11.365353 |
| 1984 | 8.50E+04 | Apple Macintosh | 11.350759 |
| 1986 | 5.38E+05 | Compaq Deskpro 386 | 13.195986 |
| 1987 | 2.33E+05 | Apple Mac II | 12.357076 |
| 1993 | 3.55E+06 | Pentium PC | 15.082176 |
| 1996 | 4.81E+07 | Pentium PC | 17.688377 |
| 1998 | 1.33E+08 | Pentium II PC | 18.708113 |
| 1999 | 7.03E+08 | Pentium III PC | 20.370867 |
| 2000 | 1.09E+08 | IBM ASCI White | 18.506858 |
| 2000 | 3.40E+08 | Power Macintosh G4/500 | 19.644456 |
| 2003 | 2.07E+09 | Power Macintosh G5 2.0 | 21.450814 |
| 2004 | 3.49E+09 | Dell Dimension 8400 | 21.973168 |
| 2005 | 6.36E+09 | Power Mac G5 Quad | 22.573294 |
| 2008 | 3.50E+10 | Dell XPS 630 | 24.278614 |
| 2008 | 2.07E+10 | Mac Pro | 23.7534 |
| 2009 | 1.63E+10 | Intel Core i7 Desktop | 23.514431 |
| 2010 | 5.32E+10 | Intel Core i7 Desktop | 24.697324 |


11. Top 500 Supercomputer Sites, http://top500.org/.

12. "Microprocessor Quick Reference Guide," Intel Research, http://www.intel.com/pressroom/kits/quickreffam.htm.

13. 1971–2000: VLSI Research Inc.

2001–2006: *The International Technology Roadmap for Semiconductors*, 2002 Update and 2004 Update, Table 7a, "Cost— Near-term Years," "DRAM cost/bit at (packaged microcents) at production."

2007–2008: *The International Technology Roadmap for Semiconductors*, 2007, Tables 7a and 7b, "Cost— Near-term Years," "Cost— Longterm Years," http://www.itrs.net/Links/2007ITRS/ExecSum2007.pdf.

2009–2022: *The International Technology Roadmap for Semiconductors*, 2009, Tables 7a and 7b, "Cost— Near-term Years," "Cost— Longterm Years," http://www.itrs.net/Links/2009ITRS/Home2009.htm.

14. Para que todos los valores en dólares fueran comparables, los precios de los ordenadores correspondientes a todos los anios han sido convertidos en dólares de anio 2000 usando los datos del *Federal Reserve Board's CPI* encontrados en: http://minneapolisfed.org/research/data/us/calc/. Por ejemplo, $1 millón en 1960 equivale a $5,8 millones en 2000, y $1 millón en 2004 equivale a $0,91 millones en 2000.

1949: http://www.cl.cam.ac.uk/UoCCL/misc/EDSAC99/statistics.html, http://www.davros.org/misc/chronology.html.

1951: Richard E. Matick, *Computer Storage Systems and Technology* (New York: John Wiley & Sons, 1977); http://inventors.about.com/library/weekly/aa062398.htm.

1955: Matick, *Computer Storage Systems and Technology*; OECD, 1968, http://members.iinet.net.au/~dgreen/timeline.html.

1960: ftp://rtfm.mit.edu/pub/usenet/alt.sys.pdp8/PDP-8_Frequently_Asked_Questions_%28posted_every_other_month%29; http://www.dbit.com/~greeng3/pdp1/pdp1.html#INTRODUCTION.

1962: ftp://rtfm.mit.edu/pub/usenet/alt.sys.pdp8/PDP-8_Frequently_Asked_Questions_%28posted_every_other_month%29.

1964: Matick, *Computer Storage Systems and Technology*; http://www.research.microsoft.com/users/gbell/craytalk; http://www.ddj.com/documents/s=1493/ddj0005hc/.

1965: Matick, *Computer Storage Systems and Technology*; http://www.fourmilab.ch/documents/univac/config1108.html; http://www.frobenius.com/univac.htm.

1968: Data General.

1969, 1970: http://www.eetimes.com/special/special_issues/millennium/milestones/whittier.html.

1974: Scientific Electronic Biological Computer Consulting (SCELBI).

1975–1996: *Byte* magazine advertisements.

1997–2000: *PC Computing* magazine advertisements.

2001: www.pricewatch.com (http://www.jc-news.com/parse.cgi?-news/pricewatch/raw/pw-010702).

2002: www.pricewatch.com (http://www.jc-news.com/parse.cgi?-news/pricewatch/raw/pw-020624).

2003: http://sharkyextreme.com/guides/WMPG/article.php/10706_2227191_2.

2004: http://www.pricewatch.com (11/17/04).

2008: http://www.pricewatch.com (10/02/08) ($16.61).

15. Dataquest/ Intel and PathfinderResearch:

| *Año* | *$* | *Log ($)* |
|---|---|---|
| 1968 | 1.00000000 | 0 |
| 1969 | 0.85000000 | −0.16252 |
| 1970 | 0.60000000 | −0.51083 |
| 1971 | 0.30000000 | −1.20397 |
| 1972 | 0.15000000 | −1.89712 |
| 1973 | 0.10000000 | −2.30259 |
| 1974 | 0.07000000 | −2.65926 |
| 1975 | 0.02800000 | −3.57555 |
| 1976 | 0.01500000 | −4.19971 |
| 1977 | 0.00800000 | −4.82831 |
| 1978 | 0.00500000 | −5.29832 |
| 1979 | 0.00200000 | −6.21461 |
| 1980 | 0.00130000 | −6.64539 |
| 1981 | 0.00082000 | −7.10621 |

| | | |
|---|---|---|
| 1982 | 0.00040000 | −7.82405 |
| 1983 | 0.00032000 | −8.04719 |
| 1984 | 0.00032000 | −8.04719 |
| 1985 | 0.00015000 | −8.80488 |
| 1986 | 0.00009000 | −9.31570 |
| 1987 | 0.00008100 | −9.42106 |
| 1988 | 0.00006000 | −9.72117 |
| 1989 | 0.00003500 | −10.2602 |
| 1990 | 0.00002000 | −10.8198 |
| 1991 | 0.00001700 | −10.9823 |
| 1992 | 0.00001000 | −11.5129 |
| 1993 | 0.00000900 | −11.6183 |
| 1994 | 0.00000800 | −11.7361 |
| 1995 | 0.00000700 | −11.8696 |
| 1996 | 0.00000500 | −12.2061 |
| 1997 | 0.00000300 | −12.7169 |
| 1998 | 0.00000140 | −13.4790 |
| 1999 | 0.00000095 | −13.8668 |
| 2000 | 0.00000080 | −14.0387 |
| 2001 | 0.00000035 | −14.8653 |
| 2002 | 0.00000026 | −15.1626 |
| 2003 | 0.00000017 | −15.5875 |
| 2004 | 0.00000012 | −15.9358 |
| 2005 | 0.000000081 | −16.3288 |
| 2006 | 0.000000063 | −16.5801 |
| 2007 | 0.000000024 | −17.5452 |
| 2008 | 0.000000016 | −17.9507 |

16. Steve Cullen, In-Stat, September 2008, www.instat.com.

| *Año* | *Mbits* | *Bits* |
|---|---|---|
| 1971 | 921.6 | 9.216E+08 |
| 1972 | 3788.8 | 3.789E+09 |
| 1973 | 8294.4 | 8.294E+09 |
| 1974 | 19865.6 | 1.987E+10 |
| 1975 | 42700.8 | 4.270E+10 |
| 1976 | 130662.4 | 1.307E+11 |

| | | |
|---|---|---|
| 1977 | 276070.4 | 2.761E+11 |
| 1978 | 663859.2 | 6.639E+11 |
| 1979 | 1438720.0 | 1.439E+12 |
| 1980 | 3172761.6 | 3.173E+12 |
| 1981 | 4512665.6 | 4.513E+12 |
| 1982 | 11520409.6 | 1.152E+13 |
| 1983 | 29648486.4 | 2.965E+13 |
| 1984 | 68418764.8 | 6.842E+13 |
| 1985 | 87518412.8 | 8.752E+13 |
| 1986 | 192407142.4 | 1.924E+14 |
| 1987 | 255608422.4 | 2.556E+14 |
| 1988 | 429404979.2 | 4.294E+14 |
| 1989 | 631957094.4 | 6.320E+14 |
| 1990 | 950593126.4 | 9.506E+14 |
| 1991 | 1546590618 | 1.547E+15 |
| 1992 | 2845638656 | 2.846E+15 |
| 1993 | 4177959322 | 4.178E+15 |
| 1994 | 7510805709 | 7.511E+15 |
| 1995 | 13010599936 | 1.301E+16 |
| 1996 | 23359078007 | 2.336E+16 |
| 1997 | 45653879161 | 4.565E+16 |
| 1998 | 85176878105 | 8.518E+16 |
| 1999 | 1.47327E+11 | 1.473E+17 |
| 2000 | 2.63636E+11 | 2.636E+17 |
| 2001 | 4.19672E+11 | 4.197E+17 |
| 2002 | 5.90009E+11 | 5.900E+17 |
| 2003 | 8.23015E+11 | 8.230E+17 |
| 2004 | 1.32133E+12 | 1.321E+18 |
| 2005 | 1.9946E+12 | 1.995E+18 |
| 2006 | 2.94507E+12 | 2.945E+18 |
| 2007 | 5.62814E+12 | 5.628E+18 |

17. "Historical Notes about the Cost of Hard Drive Storage Space," http://www.littletechshoppe.com/ns1625/winchest.html; *Byte* magazine advertisements, 1977–1998; *PC Computing* magazine advertisements, 3/1999; *Understanding Computers: Memory and Storage* (New York: Time Life, 1990); http://www.cedmagic.com/history/ibm-305-ramac.html; John C. McCallum, "Disk Drive Prices (1955–2012)," http://www.jcmit.com/diskprice.htm; IBM, "Frequently Asked Questions," http://www-03.ibm.com/ibm/history/documents/pdf/faq.pdf; IBM, "IBM 355 Disk Storage Unit," http://www-03.ibm.com/ibm/history/exhibits/storage/storage_355.html; IBM, "IBM 3380 Direct Access Storage Device," http://www.03-ibm.com/ibm/history/exhibits/storage/storage_3380.html.

18. "Without Driver or Map, Vans Go from Italy to China," *Sydney Morning Herald*, October 29, 2010, http://www.smh.com.au/technology/technology-news/without-driver-or-map-vans-go-from-italy-to-china-20101029-176ja.html.

19. KurzweilAI.net.

20. Adaptación permitida por Amiram Grinvald y Rina Hildesheim, "VSDI: A New Era in Functional Imaging of Cortical Dynamics," *Nature Reviews Neuroscience* 5 (November 2004): 874–85.

Las herramientas principales para tomar imágenes del cerebro se muestran en este diagrama. Sus capacidades vienen representadas mediante los rectángulos.

La resolución especial se refiere a la dimensión más pequeña que puede ser medida mediante cualquier técnica. La resolución temporal el tiempo o la duración de la imagen. Cada técnica tiene sus puntos fuertes. Por ejemplo, EEG (*electroencephalography*), que mide las "ondas cerebrales" (las señales eléctricas procedentes de las neuronas), puede medir ondas cerebrales muy rápidas (que ocurren en intervalos de tiempo muy cortos), pero solo pueden recibir señales cercanas a la superficie del cerebro.

Por el contrario fMRI (*functional magnetic resonance imaging*), que usa una máquina MRI especial para medir el flujo sanguíneo que reciben las neuronas (lo cual indica la actividad neuronal), puede recibir señales muchos más profundas del cerebro (y de la médula espinal) y con una mayor resolución que llega a las decenas de micras (millonésimas partes de un metro). Sin embargo, fMRI, comparado con EEG, funciona muy lentamente.

Estas son técnicas no invasivas (no se necesita ni cirugía ni medicamentos).MEG (*magnetoencephalography*) es otra técnica no invasiva. Detecta campos magnéticos generados por las neuronas. MEG y EEG pueden resolver eventos con una resolución temporal de hasta un milisegundo, y lo pueden hacer mejor que fMRI, que como mucho puede resolver eventos con una resolución de varios cientos de milisegundos. MEG también ubica con precisión áreas del sistema auditivo primario, somatosensorial y motor.

La toma de imágenes óptica cubre casi todo el rango de resoluciones espaciales y temporales, pero es invasiva. VSDI (*voltage-sensitive dyes*) es el método más sensible para medir la actividad cerebral, pero sus mediciones se limitan a las cercanías de la superficie del neocórtex de animales.

La parte expuesta del neocórtex está cubierta por una cámara sellada transparente. Después de manchar el neocórtex con un tinte apropiado a su voltaje, se ilumina y se toma una secuencia de imágenes mediante una cámara de alta velocidad. Otras técnicas ópticas usadas en el laboratorio incluyen imágenes mediante iones (normalmente iones de calcio y sodio) a sistemas por imágenes fluorescentes (imágenes confocales e imágenes multifotónicas).

Otras técnicas de laboratorio incluyen PET (*positron emission tomography*, una técnica de imágenes de la medicina nuclear que produce imágenes en 3D), 2DG (*2-deoxyglucose* o análisis de tejidos), lesiones (que involucran daños de neuronas de animales y la observación de sus efectos), *patch clamping* (para medir las corrientes de iones a través de las membranas biológicas) y el microscopio electrónico (que usa una proyección de electrones para examinar tejidos o células a una escala muy sutil). Estas técnicas también pueden integrarse con la toma de imágenes óptica.

21. Resolución espacial MRI en micras (μm), 1980–2012:

| *Año* | *Resolución en micras* | *Mención* | *URL* |
|---|---|---|---|
| 2012 | 125 | "Characterization of Cerebral White Matter Properties Using Quantitative Magnetic Resonance Imaging Stains" | http://dx.doi.org/10.1089/brain.2011.0071 |
| 2010 | 200 | "Study of Brain Anatomy with High-Field MRI: Recent Progress" | http://dx.doi.org/10.1016/j.mri.2010.02.007 |
| 2010 | 250 | "High-Resolution Phased-Array MRI of the Human Brain at 7 Tesla: Initial Experience in Multiple Sclerosis Patients" | http://dx.doi.org/10.1111/j.1552-6569.2008.00338.x |
| 1994 | 1,000 | "Mapping Human Brain Activity in Vivo" | http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1011409/ |
| 1989 | 1,700 | "Neuroimaging in Patients with Seizures of Probable Frontal Lobe Origin" | http://dx.doi.org/10.1111/j.1528-1157.1989.tb05470.x |
| 1985 | 1,700 | "A Study of the Septum Pellucidum and Corpus Callosum in Schizophrenia with MR Imaging" | http://dx.doi.org/10.1111/j.1600-0447.1985.tb02634.x |
| 1983 | 1,700 | "Clinical Effi ciency of Nuclear Magnetic Resonance Imaging" | http://radiology.rsna.org/content/146/1/123.short |
| 1980 | 5,000 | "In Vivo NMR Imaging in Medicine: The Aberdeen Approach, Both Physical and Biological [and Discussion]" | http://dx.doi.org/10.1098/rstb.1980.0071 |

22. Resolución espacial en nanómetros (nm) de las técnicas destructivas para obtener imágenes, 1983–2011:

22. Resolución espacial en nanómetros (nm) de las técnicas destructivas para obtener imágenes, 1983–2011:

| *Año* | *x-y res (nm)* | *Mención* | *URL* | *Technique* | *Notas* |
|---|---|---|---|---|---|
| 2011 | 4 | "Focused Ion Beam Milling and Scanning Electron Microscopy of Brain Tissue" | http://dx.doi .org/10.3791/2588 | Focused ion beam/ scanning electron microscope (FIB/SEM) | |
| 2011 | 4 | "Volume Electron Microscopy for Neuronal Circuit Reconstruction" | http://dx.doi.org/10.1016 /j.conb.2011.10.022 | Scanning electron microscopy (SEM) | |
| 2011 | 4 | "Volume Electron Microscopy for Neuronal Circuit Reconstruction" | http://dx.doi.org/10.1016 /j.conb.2011.10.022 | Transmission electron microscopy (TEM) | |
| 2004 | 13 | "Serial Block-Face Scanning Electron Microscopy to Reconstruct Three-Dimensional Tissue Nanostructure" | http://dx.doi.org/10.1371 / journal.pbio.0020329 | Serial block-face scanning electron microscopy (SBF-SEM) | Result quoted in http://faculty. cs.tamu.edu/choe/ftp/ publications/choe. hpc08-preprint.pdf, provided by Yoonsuck Choe. |
| 2004 | 20 | "Wet SEM: A Novel Method for Rapid Diagnosis of Brain Tumors" | http://dx.doi.org/10.1080 /01913120490515603 | "Wet" scanning electron microscopy (wet SEM) | |
| 1998 | 100 | "A Depolarizing Chloride Current Contributes to Chemoelectrical Transduction in Olfactory Sensory Neurons in Situ" | http://www.jneurosci.org/ content/18/17/6623.full | Scanning transmission electron microscope (STEM) | |
| 1994 | 2000 | "Enhanced Optical Imaging of Rat Gliomas and Tumor Margins" | http://journals.lww. com/neurosurgery/ Abstract/1994/11000/Enhanced_ Optical_Imaging_of_Rat_ Gliomas_and_Tumor.19.aspx | Enhanced optical imaging | With a spatial resolution of the optical images below 20 microns 2/pixel (22). |
| 1983 | 3000 | "3D Imaging of X-Ray Microscopy" | http://www.scipress.org/e-library/sof2/pdf/0105.PDF | Projection microscopy | See Fig. 7 in article. |

23. Resolución espacial en micras (µm) de técnicas no destructives para toma r imágenes en animales, 85–2012:

| **Año** | **Descubrimiento** | |
|---|---|---|
| 2012 | Resolución | 0.07 |
| | Mención | Sebastian Berning et al., "Nanoscopy in a Living Mouse Brain," *Science* 335, no. 6068 (February 3, 2012): 551. |
| | URL | http://dx.doi.org/10.1126/science.1215369 |

| | | |
|---|---|---|
| | Technique | Stimulated emission depletion (STED) fl uorescence nanoscopy |
| | Notas | Highest resolution achieved in vivo so far |
| 2012 | Resolución | 0.25 |
| | Mención | Sebastian Berning et al., "Nanoscopy in a Living Mouse Brain," *Science* 335, no. 6068 (February 3, 2012): 551. |
| | URL | http://dx.doi.org/10.1126/science.1215369 |
| | Technique | Confocal and multiphoton microscopy |
| 2004 | Resolución | 50 |
| | Mención | Amiram Grinvald and Rina Hildesheim, "VSDI: A New Era in Functional Imaging of Cortical Dynamics," *Nature Reviews Neuroscience* 5 (November 2004): 874–85. |
| | URL | http://dx.doi.org/10.1038/nrn1536 |
| | Technique | Imaging based on voltage-sensitive dyes (VSDI) |
| | Notas | "VSDI has provided high-resolution maps, which correspond to cortical columns in which spiking occurs, and off er a spatial resolution better than 50 µm." |
| 1996 | Resolución | 50 |
| | Mención | Dov Malonek and Amiram Grinvald, "Interactions between Electrical Activity and Cortical Microcirculation Revealed by Imaging Spectroscopy: Implications for Functional Brain Mapping," *Science* 272, no. 5261 (April 26, 1996): 551–54. |
| | URL | http://dx.doi.org/10.1126/science.272.5261.551 |
| | Technique | Imaging spectroscopy |
| | Notas | "The study of spatial relationships between individual cortical columns within a given brain area has become feasible with optical imaging based on intrinsic signals, at a spatial resolution of about 50 µm." |
| 1995 | Resolución | 50 |
| | Mención | D. H. Turnbull et al., "Ultrasound Backscatter Microscope Analysis of Early Mouse Embryonic Brain Development," *Proceedings of the National Academy of Sciences* 92, no. 6 (March 14, 1995): 2239–43. |
| | URL | http://www.pnas.org/content/92/6/2239.short |
| | Technique | Ultrasound backscatter microscopy |
| | Notas | "We demonstrate application of a real-time imaging method called ultrasound backscatter microscopy for visualizing mouse early embryonic neural tubes and hearts. This method was used to study live embryos in utero between 9.5 and 11.5 days of embryogenesis, with a spatial resolution close to 50 µm." |
| 1985 | Resolución | 500 |
| | Mención | H. S. Orbach, L. B. Cohen, and A. Grinvald, "Optical Mapping of Electrical Activity in Rat Somatosensory and Visual Cortex," *Journal of Neuroscience* 5, no. 7 (July 1, 1985): 1886–95. |
| | URL | http://www.jneurosci.org/content/5/7/1886.short |
| | Technique | Optical methods |

# Capítulo once: Objeciones

1. Paul G. Allen and Mark Greaves, "Paul Allen: The Singularity Isn't Near," *Technology Review*, October 12,

2011, http://www.technologyreview.com/blog/guest/27206/.

2. ITRS, "International Technology Roadmap for Semiconductors," http://www.itrs.net/Links/2011ITRS/Home2011.htm.

3. Ray Kurzweil, *The Singularity Is Near* (New York: Viking, 2005), chapter 2.

4. La nota 2 del escrito de Allen y Greaves "La Singularidad no está cerca" dice lo siguiente: "Nos estamos acercando a los niveles de capacidades informáticas que nos pueden permitir realizar este tipo de simulación cerebral masiva. Ordenadores pentaflop como el BlueGene/P de IBM usado en el sistema Watson ya están disponibles en el mercado. Los ordenadores exaflop están siendo diseñados. Es probable que estos sistemas pudieran llevar a cabo la computación necesaria para simular el disparo de patrones de todas las neuronas del cerebro, aunque lo harían mucho más despacio que un cerebro real".

5. Kurzweil, *The Singularity Is Near*, chapter 9, section titled "The Criticism from Software" (pp. 435–42).

6. Ibid., chapter 9.

7. Aunque no es posible determinar con precisión el contenido de la información del genoma porque la repetición de los pares de bases es muy inferior al total de datos no comprimidos. He aquí dos estrategias para calcular la información comprimida contenida en el genoma. Ambas demuestran que un rango entre 30 y 100 millones de bytes es conservadoramente alto.
   1. En términos de datos no comprimidos, existen 3 mil millones de anillos de ADN en el código genético humano, cada uno de los cuales codifica 2 bits (ya que existen 4 posibilidades por cada base de datos de ADN). Así, el genoma humano consta de alrededor de 800 millones de bytes no comprimidos. El ADN no codificante solía recibir el nombre de "ADN basura", pero ahora se sabe que juega un papel importante en la expresión génica. Sin embargo, está codificado de forma muy ineficiente. Entre otras cosas, existen enormes redundancias (por ejemplo, la secuencia llamada "ALU" se repite cientos de miles de veces), razón por la cual podemos sacar partido de los algoritmos de compresión.

      Tras la reciente explosión de los bancos de datos genéticos, existe un gran interés en la compresión de los datos genéticos. Trabajos recientes para aplicar los algoritmos de compresión de datos estándar a los datos genéticos indican que reducir los datos en un 90% (para lograr una compresión perfecta de los bits) es posible: Hisahiko Sato et al., "DNA Data Compression in the Post Genome Era," *Genome Informatics* 12 (2001): 512–14, http://www.jsbi.org/journal/GIW01/GIW01P130.pdf.

      Así, podemos comprimir el genoma a unos 80 millones de bytes sin perder información (lo cual significa que podemos reconstruir perfectamente todos los 800 millones de bytes que no están comprimidos en el genoma).

      Consideremos ahora que más del 98% del genoma no codifica proteínas. Incluso tras la comprensión estándar de datos (que elimina redundancias y usa un diccionario de consulta para las secuencias habituales), el contenido algorítmico de las regiones no codificantes parece ser bastante bajo, lo cual significa que es posible que pudiéramos codificar un algoritmo que realizara la misma función mediante menos bits. Sin embargo, como todavía estamos en el principio del proceso para aplicar la ingeniería inversa al cerebro, no podemos calcular de forma fiable esta nueva disminución basándonos en un algoritmo funcionalmente equivalente. Por tanto, uso un rango entre 30 y 100 millones de bytes de información comprimida en el genoma. La parte superior de este rango refleja solo compresión de datos y no simplificación algorítmica.

      Solo una parte (aunque mayoritaria) de esta información define el diseño del cerebro.
   2. Otra línea de razonamiento es la siguiente. Aunque el genoma humano contiene unos 3 mil millones de bases, solo un pequeño porcentaje (tal y como se señala anteriormente) codifica proteínas. Según estimaciones recientes, hay 26 000 genes que codifican proteínas. Si asumimos que esos genes tienen de media 3 000 bases de datos útiles, esto equivale aproximadamente a solo 78 millones de bases. Una base de AND requiere solo 2 bits, lo que traduce en unos 20 millones de bytes (78 millones de bases divididas por 4). En la codificación de las secuencias de proteínas de un gen, cada "palabra" (codón) de tres bases de ADN se traduce en un aminoácido. Por tanto, hay $4^3$ (64) posibles códigos de codones, cada uno consistente en tres pares de ADN. Sin embargo, hay solo 20 aminoácidos más un codón de detención (aminoácido nulo) de entre los 64. El resto de los 43 códigos se usan como sinónimos de los 21 útiles.

Mientras que se necesitan 6 bits para codificar las 64 posibles combinaciones, solo unos 4,4 ($\log_2 21$) bit son necesarios para codificar las 21 posibilidades, un ahorro de 1,6 por 6 bits (alrededor de un 27%). Estos nos lleva a una reducción que alcanza los 15 millones de bytes. Además, cierta compresión estándar basada en la repetición de secuencia es posible, aunque en esta parte del ADN codificante de proteínas la compresión que se puede hacer es mucho menos que en el llamado ADN basura, que tiene enormes redundancias. Así, es probable que esto rebaje la cifra por debajo de los 12 millones de bytes. Sin embargo, ahora tenemos que añadir información para la parte no codificante del ADN que controla la expresión génica. Aunque esta parte del ADN constituye el grueso del genoma, parece tener un nivel de información bajo en contenidos y está llena de enormes redundancias. Si estimamos que se corresponde con los aproximadamente 12 millones de bytes de ADN codificante de proteínas, de nuevo llegamos a unos 24 millones de bytes aproximadamente. Desde esta perspectiva, una estimación de entre 30 y 100 millones de bytes es conservadoramente alta.

8. Dharmendra S. Modha et al., "Cognitive Computing," *Communications of the ACM* 54, no. 8 (2011): 62–71, http://cacm.acm.org/magazines/2011/8/114944-cognitive-computing/fulltext.

9. Kurzweil, *The Singularity Is Near*, chapter 9, section titled "The Criticism from Ontology: Can a Computer Be Conscious?" (pp. 458–69).

10. Michael Denton, "Organism and Machine: The Flawed Analogy," in *Are We Spiritual Machines? Ray Kurzweil vs. the Critics of Strong AI* (Seattle: Discovery Institute, 2002).

11. Hans Moravec, *Mind Children* (Cambridge, MA: Harvard University Press, 1988).

# Epílogo

1 . "In U. S., Optimism about Future for Youth Reaches All-Time Low," Gallup Politics, May 2, 2011, http://www.gallup.com/poll/147350/optimism-future-youth-reaches-time-low.aspx.

2. James C. Riley, *Rising Life Expectancy: A Global History* (Cambridge: Cambridge University Press, 2001).

3. J. Bradford DeLong, "Estimating World GDP, One Million B. C.—Present," May 24, 1998, http://econ161.berkeley.edu/TCEH/1998_Draft/World_GDP/Estimating_World_GDP.html, and http://futurist.typepad.com/my_weblog/2007/07/economic-growth.html. Véase también Peter H. Diamandis and Steven Kotler, *Abundance: The Future Is Better Than You Think* (New York: Free Press, 2012).

4. Martine Rothblatt, *Transexualidad a Transhumanidad (Transgender to Transhuman)* (edición privada, 2011). Aquí explica cómo lo más probable es que los "transhumanos" sean aceptados siguiendo una rápida trayectoria que será similar; por ejemplo, las mentes no biológicas pero convincentemente conscientes expuestas en el capítulo 9.

5. El siguiente extracto de *La Singularidad está cerca*, cap. 3 (pp 145–147), de Ray Kurzweil (Berlín: Lola Books, 2012), expone los límites de la computación basándose en las leyes de la física:

> […] los límites últimos de los ordenadores son profundamente elevados. Basándose en los trabajos del profesor de la *University of California at Berkeley* Hans Bremermann y del teórico de la nanotecnología Robert Freitas, el profesor del MIT Seth Lloyd ha calculado la capacidad de computación máxima, de acuerdo con las leyes de la física conocidas, de un ordenador que pese un kilogramo y ocupe un volumen de un litro (más o menos el tamaño y peso de un ordenador portátil pequeño). A esto lo ha llamado "el ordenador portátil primordial".
>
> La cantidad potencial de computación crece con la energía disponible. Así, la relación entre la energía y la capacidad de computación se puede entender de la siguiente manera: la energía contenida en una cantidad determinada de materia es la energía asociada con cada átomo (y con cada partícula subatómica), de manera que, cuantos más átomos, más energía. Tal y como ha quedado dicho, potencialmente cada

átomo puede ser utilizado para realizar computación, de manera que, cuantos más átomos, más computación. Así mismo, la energía de cada átomo o partícula crece con la frecuencia de su movimiento: cuanto más movimiento, más energía; y además la misma relación se establece para el potencial de computación: a mayor frecuencia de movimiento, mayor es la computación que cada componente (por ejemplo, un átomo) puede realizar. (Esto lo podemos observar en los chips actuales: a mayor frecuencia del chip, mayor es su velocidad de computación). De esto se sigue que existe una relación directamente proporcional entre la energía de un objeto y su potencial para realizar computación.

La energía potencial contenida en un kilo de materia es muy grande, tal y como nos lo indica la ecuación de Einstein $E=mc^2$ (el cuadrado de la velocidad de la luz es un número muy elevado: aproximadamente $10^{17}$ metros$^2$/segundo$^2$). A su vez, el potencial de la materia para realizar computación viene definido por un número muy pequeño, la constante de Planck: $6{,}6 \cdot 10^{-34}$ julios por segundo (el julio es una medida de energía). Esta es la escala más pequeña a la que podemos aplicar energía para computar. El límite teórico de un objeto para realizar computación lo obtenemos dividiendo la energía total (la energía media de cada átomo o partícula multiplicada por el número de dichas partículas) por la constante de Planck.

Lloyd demuestra que la capacidad potencial para realizar computación contenida en un kilogramo de materia es igual a pi (p) multiplicado por la energía dividida por la constante de Planck. Dado que la energía es un número tan elevado y la constante de Planck es tan pequeña, esta ecuación da como resultado un número extremadamente grande: unos $5 \cdot 10^{50}$ operaciones por segundo.

Si esta cantidad la contrastamos con la estimación más conservadora sobre la capacidad de los cerebros humanos ($10^{19}$ cps y $10^{10}$ humanos), obtenemos el equivalente a unos cinco millones de billones de civilizaciones humanas. Si tomamos la cantidad de $10^{16}$ cps, que yo creo que será suficiente para emular funcionalmente la inteligencia humana, el ordenador portátil primordial funcionaría con una capacidad cerebral equivalente a cinco billones de billones de civilizaciones humanas. Un ordenador portátil de estas características podría realizar el equivalente a todo el pensamiento humano de los últimos diez mil años (lo que es igual a diez mil millones de cerebros humanos funcionado durante diez mil años) en una diez milésima de nanosegundo.

Sin embargo, existen nuevas salvedades. La conversión de toda la masa de nuestro ordenador portátil de 2,2 libras en energía, es lo que esencialmente tiene lugar en una explosión termonuclear. Por supuesto, no queremos que el portátil explote, sino que se mantenga dentro de su dimensión de un litro. De manera que esto requerirá cuanto menos un embalaje muy cuidadoso.

Mediante el análisis de la entropía máxima (niveles de libertad representados por el estado de todas las partículas) contenida en un dispositivo de estas características, Lloyd demuestra que un ordenador así teóricamente tendría una capacidad de memoria de $10^{31}$ bits. Es difícil imaginar tecnologías que consigan llegar al límite de estas capacidades. Sin embargo, podemos fácilmente prever las tecnologías que se acerquen razonablemente a este límite. Tal y como indica el experimento de la Universidad de Oklahoma, ya hemos demostrado la capacidad de almacenar al menos cincuenta bits de información por átomo (aunque de momento solo en una pequeña cantidad de átomos), de manera que eventualmente sería posible almacenar $10^{27}$ bits de memoria en los $10^{25}$ átomos contenidos en un kilo de materia.

No obstante, dado que muchas propiedades de cada átomo podrían ser utilizadas para almacenar información (como por ejemplo la posición exacta, el *spin* y el estado cuántico de todas sus partículas), es posible que podamos conseguir una marca algo superior a los $10^{27}$ bits. El neurocientífico Anders Sandberg estima que la capacidad potencial de almacenamiento contenida en un átomo de hidrógeno es de unos cuatro millones de bits. Sin embargo, estas densidades no han sido demostradas todavía, por lo que utilizaremos una estimación más conservadora.

Tal y como se ha dicho anteriormente, se podrían conseguir $10^{42}$ cálculos por segundo sin producir ninguna cantidad de calor apreciable. Mediante el pleno despliegue de las técnicas de computación reversible, el uso de diseños que generen bajos niveles de error y permitiendo razonables cantidades de disipación de energía, deberíamos acabar por conseguir entre $10^{42}$ y $10^{50}$ cálculos por segundo.

El diseño del terreno entre estos dos límites es complejo y el examinar las cuestiones técnicas surgidas de pasar desde $10^{42}$ hasta $10^{50}$ sobrepasa el marco de este capítulo. Sin embargo, debemos tener en cuenta que, basándonos en varios aspectos prácticos, la manera en la que se llevará a cabo este proceso no

es empezando por el límite final de $10^{50}$ y luego trabajar hacia atrás. Más bien ocurrirá que la tecnología continuará incrementándose mediante el uso de los últimos avances para progresar hasta el siguiente nivel. Así, una vez que consigamos una civilización con $10^{42}$ cps (para cada 2,2 libras de peso), los científicos e ingenieros de ese día utilizarán su vasta inteligencia (que en su parte fundamental no será biológica) para hallar la manera de llegar hasta $10^{43}$, luego hasta $10^{44}$, y así sucesivamente. Mi previsión es que llegaremos a estar muy cerca de estos límites finales.

Incluso a un nivel de $10^{42}$ cps, un "ordenador portátil primordial" de 2,2 libras de peso podría realizar el equivalente a todo el pensamiento humano de los últimos diez mil años en diez microsegundos (si damos por buena la estimación de diez mil millones de cerebros humanos durante un periodo de diez mil años). Si observamos con atención la gráfica de la página 77, "el crecimiento exponencial de la capacidad de computación", vemos que esta cantidad de computación se estima que sea posible por un precio de mil dólares hacia el año 2080.

# Índice

**A**
Ackerman, Diane 171
actividad mental 219
actividad neuronal 219, 297
Adams, Douglas 154
adrenalina 101
*Age of Intelligent Machines, The* ix, 3, 158, 242–43
*Age of Spiritual Machines, The* ix, 3, 243, 253
Aiken, Howard 180
ajedrez x, 6, 35, 154, 158, 173, 243, 255
Alexander, Richard D. 213
algoritmo 5, 22, 68, 78, 81, 83–85, 107, 142–43, 146, 155, 159, 165, 167, 173, 177, 183, 185–86, 258, 261, 279–80, 282–85, 303
algoritmo evolutivo 84
algoritmo genético 142–44, 155, 165, 167, 284
algoritmo neocortical 83, 146, 185
Allen, Paul 253, 302
Allman, John M. 171
almacenamiento de patrones 60
Alzheimer 97
ambiente simulado 142
amígdala 67, 101, 103, 105, 108, 120
amor xvii, 103, 105–06, 110, 112–15, 154, 210, 212
análisis estadístico 6, 162
ancho de banda de la información 102
ancho de banda del neocórtex 102
animales 2, 12, 61, 73, 89, 96, 99, 117–20, 155, 200, 202–04, 235–36, 248, 250–51, 267, 274, 297, 301
aprendizaje 47, 58–59, 61, 72–75, 79, 81, 85, 98, 114, 117–18, 120–23, 137, 141–43, 148–49, 155–57, 159–60, 167, 183, 186, 188, 257, 260
aptitud 107
Arendt, Hannah 193

argipresina 113–14
Aristóteles 228–29
arquitectura del cerebro 186
arquitectura del ordenador 179, 188
autoasociación 56, 128
autómata celular 224–25, 227
autoorganización 84, 149, 251
avatar 194, 200
axones 32, 38–39, 44–45, 52–53, 62–63, 75, 77, 85, 95, 101, 109, 144, 165–66, 182, 215

**B**

Babbage, Charles 181
Bainbridge, David 171
Barrow Neurological Institute 94
Berger, Theodore 97
Berners-Lee, Tim 165
Bierce, Ambrose 62
biología 2, 4, 6–7, 11–12, 14, 34, 84, 91, 115, 124, 145, 147, 204, 239, 262
biomedicina 237
Blackmore, Susan 202
Blakeslee, Sandra 68, 277
Blue Brain 59, 75, 120–23
Bode, Stella de 214, 288
bomba atómica 19
bonobos 106
Boyden, Ed 121
Brodsky, Joseph 191
Burns, Eric A. 108
Butler, Samuel 58, 191, 213, 235–36
Byron, Ada 181, 182

**C**

capacidad de autoreflexión 192
capacidad de reconocimiento 25
capacidades del cerebro 184
capacidad para resolver problemas 265
capa IV 32, 33
capa V 76
capa V1 78, 82, 95

capa V2 78, 82
capa VI 32, 95
Carnegie Mellon University 173
Carroll, Lewis 105
células biológicas 233
células ciliadas internas 133
células de Purkinje 98
células en huso 106
células fusiformes 105–06
células ganglionares 91
células nerviosas 84
células sanguíneas 5, 230–32, 267
centro lingüístico 215, 217
cerebelo 7, 97–99, 120, 236
cerebro antiguo 59, 67, 85, 89, 98–99, 101–02, 120, 169, 277
cerebro biológico xi, 47, 119, 122, 124, 166, 168–70, 172, 234
cerebro digital 119, 166–67, 170
cerebro humano x–xiii, 4–5, 7–8, 19, 32, 37, 61, 72, 75, 79, 81, 89, 101–03, 105, 120, 122, 124–25, 147, 157, 160, 162–64, 166, 168–69, 171, 173, 176–77, 182, 184–87, 194, 198, 247, 255, 258–62, 268
cerebro simulado 122–23
cerebros reales 20, 60, 124, 144, 302
Chalmers, David 193, 208, 229
chatbot 154, 221
chimpancés 3, 37, 106
Chomsky, Noam 52, 151, 274
Church, Alonzo 177
circuitos biológicos 48, 118
circuitos corticales 144–45
circuitos corticales biológicos 144
circuitos digitales 186–88
circuitos electrónicos 118
circuitos neocorticales 101, 187
circuitos neuronales 176
civilización humano-máquina xii, 269
clon genético 230
clon mental 230
coches autoconducidos 152, 260
coches autopilotados 6

Cockburn, David 204, 287
cóclea 92, 130, 133, 231
código genético 9, 33, 142, 284–85, 303
colesterol 101–02
columna cortical 33, 35, 76
columna neocortical 85, 120, 121
complejidad xi, xiii, 3–4, 8–9, 10, 12, 61, 64, 136, 149, 169, 173, 189, 194, 212, 221, 255–56, 258–59, 266
comportamiento animal 117
comportamiento aprendido 97, 117
comprensión del cerebro 34, 251
computación analógica 260
computación cuántica 198–99, 260–61
computación digital 184
computación en tres dimensiones 241, 255
computación molecular 242
*Computer and the Brain, The* (von Neumann) 182
conciencia xi, 22, 51, 231, 235, 271
conexiones físicas 61
conexiones interneuronales 186, 248, 259, 261, 282
conexiones neocorticales 80, 232
conexiones neuronales 38, 85, 166, 247
conocimiento jerárquico 118
conocimiento profesional 36, 161
conocimiento transcendente 115
consciencia 24, 71, 94, 100, 120, 185, 191–203, 205, 208, 210–13, 215–23, 229–30, 235, 261, 288
control del movimiento 32
copias de seguridad 119, 166, 234
corazón 58–59, 95, 105, 113
córtex auditivo 7, 44, 124
córtex cerebral 7, 81, 105
córtex humano 257
córtex motor 32, 218, 219
córtex simulado 145
córtex visual 7, 78, 81, 82, 184
Costanza, George 71
Craig, Arthur 94
creatividad 103, 108, 111, 212, 276

crecimiento exponencial 3, 4, 37, 111, 237, 239, 242, 255, 259, 307
Crick, Francis 14
cuantificación vectorial 130, 133, 136, 139, 141
cuerpo xvii, 66, 92–93, 97, 119, 182, 206, 211, 215–16, 230–32, 257, 272
cuerpo biológico 119, 267
cultura humana 58
Curtiss, Susan 214, 288
curva en S 241
*Cybernetics* (Wiener) 110

**D**

Dalai Lama 105
Dalrymple, David xviii, 119, 279
DARPA 154, 156
Darwin, Charles 11–14, 19, 47, 109, 111, 169, 273
datos digitales 175–76
datos genéticos 238–39, 303
datos sensoriales 57, 93, 95
década de 1930 178
década de 1940 171
década de 1950 99, 128, 185, 241, 255
década de 1960 21, 128, 241
década de 1970 130, 147, 159
década de 1980 3, 68–69, 118, 130, 136, 138–39, 147, 153, 158, 166, 242
década de 1990 3, 153, 241–43
década de 2020 122–23, 184, 255
década de 2030 201, 231
década de 2040 123
Deep Blue x, xiii, 35, 158
DeMille, Cecil B. 108
dendritas 38–39, 45, 62–63, 77, 85, 106, 144, 166, 182–83, 187, 232
Dennett, Daniel 196, 219, 222
Denton, Michael 262, 304
derechos de las máquinas 200
Descartes, René 208, 211, 228
determinismo 220–22
Diamandis, Peter ix, xviii, 266
Diamond, Marian 21
Dickens, Charles 161
Dickinson, Emily 1

Diógenes Laercio 233
diseño neocortical 10
diversidad xi, 9, 12, 14, 60, 134
doble hélice 14
dopamina 100–02, 112–13
Dostoevsky, Fyodor 191
Dragon Go! 155, 157, 286
Drave, Scott 143

**E**
Eckert, J. Presper 180
educación 21, 23, 85, 107, 167, 170
EE.UU. 11, 21, 162, 197, 237
Einstein, Albert 10, 15–21, 31, 56, 67, 109, 111–12, 177, 273–74, 305
eliminación de la pobreza 266
*El Origen de las Especies* (Darwin) 12–13
Emerson, Ralph Waldo 11
emulación cerebral completa 125
emulación del comportamiento humano 102
emulación tecnológica del neocórtex 61
entidades futuras no biológicas 204
EPAM (elementary perceiver and memorizer) 34
escaneado 123, 230, 232, 234
escaneo del cerebro 7, 124, 248, 259
escaneo no invasivo 124
especie humana 53, 263
esperanza de vida 183, 232, 237, 265
estrógeno 113
estructuras neocorticales 65, 202
evolución biológica xi, xiii, 3, 7, 13, 58, 71, 73–74, 84, 99, 107, 141, 143, 145–46, 169, 242, 268
evolución de la inteligencia 168, 173
evolución del neocórtex 118
evolución de patrones 118
evolución simulada 142–43, 145
experiencias sensoriales 65, 96
experimento de Michelson-Morley 15, 109
experimento mental x, 5, 11, 12, 14, 16–17, 19–23, 29, 67, 85–86, 108–09, 112, 162, 177, 191, 193–95, 200, 206, 216–17, 220, 222, 230–32, 261, 273, 287
expertos humanos 47, 140–41

**F**

Facebook xii, 149, 241, 290
feedback 128
Feldman, Daniel E. 83, 277
Felleman, J. 81, 276
fenómenos bioquímicos 113
fenómenos físicos 113
filosofía 105, 148, 192, 196, 209–10, 224, 228
física 1–2, 7, 11, 14, 20, 33–34, 45, 47, 73, 77, 85, 95, 164, 166, 208, 210–11, 235, 242, 254, 268–69, 272, 305
flujo de conciencia 51
flujo de datos 49
flujo de información 54
fonemas 46–47, 57–58, 60, 68–69, 130, 132, 140, 146, 167
Forest, Craig 121
formas de inteligencia extrañas 204
fotón 17, 18
FPGA (Field Programmable Gate Array) 78
Franklin, Rosalind 14
Freud, Sigmund 62, 67, 275
Fried, Itzhak 218
Friston, K. J. 71
funciones vitales 83

**G**

García Márquez, Gabriel 3, 271
Gazzaniga, Michael 215–17, 223, 288
genoma 4, 85, 98, 149, 238–39, 257–59, 303–04
geología 11–12
George, Dileep xvii–xviii, 38, 68, 149, 286
Georgia Tech 121–22
Ginet, Carl 222
giro fusiforme 82, 84, 107
glándula pituitaria 101
glóbulos blancos 231
Gödel, Kurt 178
Good, Irvin J. 268
Google ix, xii, 6, 8, 68, 124, 152, 154, 156, 173, 247, 260, 267
gorilas 106
Gran Cañón del Colorado 11

Greaves, Mark 253, 302
Grecia 11
Grötschel, Martin 256
Guerra Mundial, Primera 266

**H**
habitación china 162, 261
habla humana 48, 68–69, 87, 130, 132, 137–38, 140, 153, 167
Hameroff, Stuart 197, 260, 287
Harnad, Stevan 253
*Harry Potter y el misterio del príncipe* (Rowling) 112
*Harry Potter y el prisionero de Azkaban* (Rowling) 117
Harvard 77, 119, 124, 288, 304
Hasson, Uri 81, 276
Havemann, Joel 21
Hawkins, Jeff 38, 68, 149
Hebb, Donald O. 75
hemisferio del cerebro 96, 101, 214, 216
hemisferio derecho 106, 215–17
hemisferio izquierdo 214–17
hemisferios cerebrales 215–16
hemisferoctomías 214
hipertensión 102
hipocampo 59–60, 96–97, 120
Hobbes, Thomas 265
Hock, Dee 108
homínidos 98, 106
homo sapiens 2, 32, 37, 74, 268
Horwitz, B. 71
Hubel, David H. 31
humanidad xiii, 108, 112, 182, 185, 266, 268
humano biológico 170, 185, 203, 205, 213
Hume, David 223, 288
humor 55, 214, 218, 228

**I**
IA fuerte 260
IBM x, xii–xiii, 6, 21, 102, 123, 157–58, 160–61, 182, 185–86, 256, 258, 287, 290–92, 296, 302
IBM BlueGene/P 123

identidad xi, 9–10, 78, 196, 228–34, 288, 289
impresiones sensoriales 197, 210–11
inconsciencia 95
información auditiva 92, 108, 260
información epigenética 257
información limitada 209
información sensorial 50, 94–96, 153
ingeniería inversa x–xi, 4–5, 7, 34, 239, 251, 258–59, 262, 267, 303
inputs nerviosos 92
inputs sensoriales 55
inputs unidimensionales 60
inspiración biológica 117, 260, 279
Institutos Nacionales de Salud 124
insulina 34, 258
inteligencia artificial (IA) ix–xiii, xvii, 6–7, 34, 47–48, 86–87, 91, 107, 123, 128, 130, 147–48, 150–53, 162, 181–82, 186, 251, 256, 258, 260, 262
inteligencia biológica 118, 164
inteligencia del mundo biológico 204
inteligencia emocional 106, 185, 193
inteligencia expandida 111
inteligencia humana x, xiii–xiv, 1, 5, 20, 59, 151–52, 158, 164, 185, 253, 269, 306
inteligencia jerárquica 164
inteligencia no biológica 204, 267
inteligencias futuras 201
International Technology Roadmap for Semiconductors 255, 292, 302
intuición humana 253
*Investigaciones Filosóficas* (Wittgenstein) 210
isleta pancreática 34, 258

**J**

James, William 71, 94
Jeffers, Susan 99
Jennings, Ken 151, 157
Jeopardy! 6, 102, 151, 153, 157–62, 164, 169, 221, 256–57
jerarquía 10, 28, 31, 44–45, 48–50, 53–55, 63–65, 68, 79, 81, 85, 118, 132, 136, 137–41, 145, 155–57, 163, 165, 167, 169, 186, 207
jerarquía de patrones 28, 31, 44, 64, 68, 137
jerarquización 28
Joyce, James 51
juicio moral 106

**K**

Kasparov, Garry x, 35, 158
King, Michael Patrick 112
Kodandaramaiah, Suhasa 121
Koene, Randal xviii, 84, 277
Koltsov, Nikolai 14
Kotler, Steven 266, 305
KurzweilAI.net xvii, 154, 296
Kurzweil Applied Intelligence 138
Kurzweil Computer Products 6, 118, 146
Kurzweil Voice 153

**L**

Larson, Gary 265
Leibniz, Gottfried Wilhelm 31, 213
Lenat, Douglas 155
lenguaje 2–3, 6, 23, 32, 34, 37–39, 48, 51–53, 57, 62, 64–66, 68, 82–83, 87, 110, 123, 132, 137–40, 146–48, 150, 152–62, 164, 166–67, 173, 188, 203, 209–10, 214, 251, 255–57, 261, 275
lenguaje natural 6, 48, 68, 87, 139, 146, 150, 152–61, 166–67, 251, 255, 257
lenguajes informáticos 148
lenguajes máquina modernos 181
Lewis, Al 89
ley de Cooper 239
ley de los rendimientos acelerados (LOAR) 3–5, 7, 37, 119, 235–37, 242–43, 253–55, 267–69, 289
ley de Moore 237, 241, 254–55
leyes de la termodinámica 34, 254
leyes naturales 177, 261
Libet, Benjamin 218, 220, 222, 288
libre albedrío xi, 10, 191, 196, 213, 217, 219–24, 227–29
LISP (LISt Processor) 147–49, 155, 159
lóbulo temporal medio 96
Lois, George 108
Lyell, Charles 12, 109, 169

**M**

mamíferos xi, 2, 32, 74–75, 89, 100–01, 106, 118, 124, 127, 166, 265, 274
Mandelbrot, conjunto 9–10
máquina artificial 193
máquina biológica 192
máquina de Turing 177–79, 183, 198
máquina inteligente 191, 262
máquina von Neumann 179–81, 183
Marconi, Guglielmo 239
Mark 1 Perceptron 126–27, 129
Márkov, Andrei Andreyevich 138
Markram, Henry 75, 120, 276, 279
Massachusetts General Hospital 77, 124
Mauchly, John 180
Maudsley, Henry 214, 288
Maxwell, James Clerk 17
Maxwell, Robert 214
McClelland, Shearwood 214
McGinn, Colin 192
Mead, Carver 185, 287
mecánica cuántica 208–09, 223–24
mecanismos analógicos 184
mecanismos del cerebro 184
mecanismos digitales 184
mecanismos hormonales 114
medicina 35, 161, 237, 239, 297
médula espinal 32, 93, 297
memoria xiii, 5, 23–25, 53, 57, 59, 68, 89, 96, 98, 112, 121, 149–51, 165–67, 175, 176–77, 179–81, 183–84, 186–87, 197, 207, 214, 231, 245–46, 255–57, 306
memoria jerárquica 96
Menabrea, Luigi 181
mente ix–x, xvii, xix, 5, 7–8, 10–11, 24, 26, 28, 31, 37, 45, 51–52, 64–65, 71, 74–75, 77, 89, 94, 107–08, 111–12, 136–37, 164–65, 171, 185, 191, 194, 196, 211, 217, 253, 274–76, 286–87
metáforas 6, 11, 16, 57, 59, 108–12, 114, 119, 158, 168–69, 256
método autoorganizativo 138, 141, 143, 146, 148, 164, 257
métodos funcionales 260
Michelson, Albert 15
microtúbulos 197, 199, 260–261

microtus montanus 113–14
microtus ochrogaster 113
miedo 67, 99–102, 108, 119, 169, 204
Miescher, Friedrich 13, 273
minicolumnas 33
Minsky, Marvin xi, xviii, 58, 128–29, 191, 217, 288
MIT (Massachusetts Institute of Technology) ix, xi, 9, 79, 96, 121–22, 128, 152, 267, 274, 279, 305
mitocondria 14
modelo algorítmico 123
modelo de neurona 182
modelos jerárquicos neocorticales autoorganizativos 150
modelos jerárquicos ocultos de Márkov (HHMMs) 48, 63, 68, 70, 138–41, 143–48, 150, 154–56, 160, 166, 186, 256–57
modelos matemáticos 99
modelos moleculares 123
modelos ocultos de Márkov 63, 136–41, 144, 166–67
modelo unificado del neocórtex 20
Modha, Dharmendra 123, 185, 258
Money, William 113
Moore, Gordon 237
Moravec, Hans 186, 262, 304
Morley, Edward 15
Moskovitz, Dustin 149
Mountcastle, Vernon 33, 90
Mozart, Wolfgang Amadeus 106
Muckli, Lars 214, 288
mundo biológico 145, 204, 234
mundo físico 19, 194, 208, 211–12
mundo no biológico 234
mutación 100, 142
mutación genética 100

**N**

Naciones Unidas 156
nanobots 267
National Institutes of Health 77, 79
National Science Foundation 77
naturaleza jerárquica 32, 52, 81, 89, 138, 152, 156
neocórtex adicional 118, 164

neocórtex artificial 118, 169
neocórtex biológico 61, 71, 119, 145, 150, 166, 186, 276
neocórtex cerebral 7
neocórtex digital x, 117–19, 166, 279
neocórtex humano 32, 34, 36, 58–59, 89, 138, 188, 221
neocórtex no biológico 102, 111–12
neocórtex simulado 121, 138, 141, 185
neocórtex sintético 37
neocórtex visual 79, 81, 123
nervio auditivo 124, 133, 231
nervio óptico 50, 90–91, 95–96
Neumann, John von 171, 178–179, 224, 286–87
neurociencia xii, 5, 7, 32–33, 74, 81, 182–83
neurocomputación 258
neuronas biológicas 130, 267
neuronas de la lámina 1 93
neuronas en huso 105
neuronas motoras 83
Newell, Allen 173
Newton, Isaac 90
Nick Bostrom 124–26, 211, 279
Nietzsche, Friedrich 112
niveles conceptuales 49, 59, 61, 157, 163, 165, 168, 173, 268
niveles hormonales 102
niveles jerárquicos 81, 111, 167
nivel molecular 34, 119, 120
nivel óptimo 61
noradrenalina 101, 112–13
Nuance 6, 102, 118, 146, 153–55, 160–61, 286
núcleo accumbens 99–100
núcleo celular 13
núcleo geniculado lateral 95
núcleo ventromedial posterior (VMpo) 94–95

**O**

oído absoluto 107–08
ojos 28, 49, 57, 59, 69, 78, 90, 108, 136, 155, 206, 213, 259, 271–72
Oluseun, Oluseyi 195
*On Intelligence* (Hawkins, Blakeslee) 68–69, 149
optimización de los parámetros 142

ordenadores analógicos 174–175
ordenadores digitales 174–75, 260
ordenador moderno 179, 184
organismo simulado 142
organización columnar del neocórtex 33
organización del cerebro 188
organización masivamente paralela del cerebro 184
oxitocina 113–14

**P**

páncreas 34, 258
Papert, Seymour 128–29
paralelismo del cerebro 184, 187
parámetros de dios 141
Parker, Sean 149
Párkinson 231–32
Pascal, Blaise 112
patrones básicos 68
patrones cerebrales 137
patrones corticales 166, 215
patrones importantes 53
patrones mentales 83
patrones neocorticales 49, 64–65, 84–85, 95, 109, 114, 123, 186, 231
patrones repetidos 36, 60
patrones sensoriales 53, 58, 78, 85
patrones simples 50, 52, 60
Pavlov, Ivan Petrovich 206, 287
Penrose, Roger 197, 260
pensamiento directo 65, 95
pensamiento exponencial 236
pensamiento humano 7, 19–20, 22, 188, 199, 306–07
pensamiento indirecto 64
pensamiento jerárquico 99, 112, 169, 221, 274
pensamiento racional 32, 211, 275
pensamientos emocionales 105
percepción sensorial 32
*Perceptrons* (Minsky, Papert) 128–30
persona consciente 200–03, 209, 212
perspectiva occidental 208
perspectiva oriental 208, 212

piedra roseta 156
piel 50, 57, 92, 136, 205, 272
Pinker, Steven 71, 266, 276
placer ix, 99–100, 102, 114, 169, 171
planeta 4, 11, 124, 265–66
plantas 61, 73
plasticidad del cerebro 214
plasticidad del neocórtex 83, 215
plasticidad neocortical 83
Platón 202, 211, 220, 229
Poggio, Tomaso xviii, 79, 152
Portman, Natalie 27
posición panprotopsíquica 203
positivismo lógico 209–10
potencia sináptica 75, 85, 101, 127
predicciones lineales 97, 236
Premio Nobel 14, 273
presuposiciones filosóficas 196, 200, 202
Princeton 81, 278
*Principia Mathematica* (Russell, Whitehead) 173
principio antrópico 1
principio de la vida 61
problema cuerpo-mente 211
problema de la invarianza 130, 132
procesadores de patrones 37, 50, 148
procesamiento de la información 92, 102, 197, 199
procesamiento jerárquico 147
procesamiento neocortical 81
procesamiento neuronal 183
procesamiento paralelo 187
procesamiento simultáneo de información 184
procesamiento visual 33, 152, 186
procesos estadísticos jerárquicos 162
procesos lógicos 35
productos químicos 100, 102, 113
programación lineal 60, 167, 256
proyecto Cyc 155, 157
puertas lógicas 176

**Q**

Quinlan, Karen Ann 95

**R**
radiómetro de Crookes 17–18
Ramachandran, Vilayanur Subramanian “Rama” 219
ratón 83–84, 111, 113–14, 124, 171
rayos X 14
reacciones emocionales 106, 193, 200, 203, 205
reconocedor del habla 133
reconocedor de patrones 9, 33, 35–40, 44–49, 52–54, 56, 58, 60–63, 69, 78, 85, 95, 109, 114, 119, 136, 147–48, 165–67, 186–87, 253, 261, 268
reconocedor jerárquico de patrones 251
reconocedor neocortical de patrones 37, 188
reconocimiento autoasociativo 165
reconocimiento del habla 46, 48, 50, 68, 92, 110–11, 118, 123, 133, 136–37, 139–40, 143–46, 154, 166–67, 188, 255, 260
reconocimiento de patrones 5, 7, 10, 31, 37, 40, 45, 47–48, 50, 57, 60, 62–63, 68–69, 74–79, 84–86, 98, 101, 107, 109–10, 130, 136–37, 141, 149–50, 164–67, 186, 262, 274, 280
reconocimientos sensoriales 55
reconocimiento visual 50
red neuronal 75, 127–28, 130, 138, 148, 279–84
reducción de datos 91, 133
redundancia 8, 36–38, 45, 53, 56–57, 60–61, 86, 167, 176, 187, 214, 251, 258
regiones cerebrales 59, 82, 105, 247, 274
regiones corticales 55, 146
regiones neocorticales 82, 85, 106
región sensorial 55
regla 110 226–27
regla 222 225–26
regla de oro 170
reglas culturales 67
reglas probabilísticas 140
relaciones sexuales 100, 114
reproducción sexual 113, 142
resolver un problema 65, 111, 118, 143, 265
resonancia magnética funcional (fMRI) 105, 297
ritmo cardiaco 101
robot 120–22, 150, 173, 194, 200, 204
robótica 121

Rosenblatt, Frank 126, 129, 182
Roska, Boton 91
Rothblatt, Martine xvii–xviii, 266, 288, 305
Rowling, J. K. 117
Russell, Bertrand 99, 173, 209
Rutter, Brad 157

**S**
salud 102, 195
Sandberg, Anders 124–26, 279, 306
Schopenhauer, Arthur 223, 228, 288
Searle, John 162, 192, 212, 261
Segunda Guerra Mundial 266
señales sensoriales 105
sentido común 36, 155, 173
sentimientos xiii, 100–02, 112–13, 192, 201, 204, 275
ser humano 14, 127, 194, 204–05, 221, 235
serotonina 100–02, 113
Seung, Sebastian 9, 273
Shakespeare 35, 109, 200
Shannon, Claude 175
Shashua, Amnon 152
Shaw, J. C. 173
siglo III a. C. 233
siglo XIII 220
siglo XIX 11, 16, 19, 31, 237
siglo XX 14, 19, 171, 181, 237
siglo XXI 237
siglo XXII 269
símbolo 2–3, 9, 108–109, 147, 162, 195, 233, 262
Simon, Herbert A. 34, 172, 274
simulación cerebral 120–21, 123, 302
simulación de la evolución 107
simulación de moléculas 124
simulación informática 119
simulación molecular 123
simular el cerebro xi, 7, 75, 119–20, 124–25
*Singularidad está cerca, La* 3–4, 186, 237, 239, 242–43, 253–56, 258, 261, 272, 305
Siri 6, 68, 111, 118, 146–47, 154, 157, 161, 163

sistema endocrino 230
sistema nervioso 101, 119, 149
sistemas autoorganizativos 144, 167
sistemas inteligentes 155, 251, 260
sistemas jerárquicos 150, 159, 256
sistemas no biológicos 233–34
Skinner, B. F. 11
Smullyan, Raymond 229, 288
sociedades humanas 102
*Society of Mind, The* (Minsky) 58, 191, 288
software del futuro 203
solución matemática 60, 112
Sperry, Roger W. 208, 288
Stanford 151, 220, 286
sueños 20, 51, 62, 65–68, 171, 275–76
superordenador 123, 158, 184, 186
supervivencia 2, 59, 72–73, 99, 169, 171, 236, 265, 286
sustancias radioactivas 19
sustrato biológico x, 201, 234
Sutherland, Stuart 202, 287
Szent-Györgyi, Albert 89

**T**

tálamo 32–33, 59, 93–96
Taylor, J. G. 71
tecnología inteligente 266–67
tecnologías de la información 3, 236–37, 239, 241–43, 254, 266
tecnologías genéticas 239
TED 120
Tegmark, Max 199
teorema de incompletud de Gödel 198
teoremas de Turing 179, 198
teoría de la mente 5, 7, 10, 31, 37, 45, 74–75, 107, 164, 274
teoría de la mente basada en el reconocimiento de patrones (PRTM) x, 5, 45, 68, 74, 81, 87, 164, 207
tesis Church-Turing 177
tesis Hameroff-Penrose 199
test de Turing x, xiii, 153, 161–62, 170, 182, 203–04, 221, 261–62, 286
testosterona 113
Thiel, Peter 149

Thrun, Sebastian 151
Tierra 12, 15–16, 58, 227, 235, 269
torrente sanguíneo 267
*Tractatus Logico-Philosophicus* (Wittgenstein) 209–210
transcendencia 211–12
tronco del encéfalo 32, 93, 98
Turing, Alan x, 117, 152, 177, 182, 268

**U**
UIMA (Unstructured Information Management Architecture) 159–60
Unión Soviética x, 14, 237, 243
universalidad de la computación 22, 173, 176, 180, 183
Universidad Cornell 126
Universidad de Oxford 124
Universidad de Texas 81
Universidad de Yale 182
University College de Londres 113
University of California at Berkeley 83, 91, 277, 305
University of California at Davis 218
University of California at Los Angeles 124
University of Minnesota 124
University of Southern California 97
universo xi, xiv, 1–2, 5, 8, 21, 31, 90, 154, 225, 227, 269, 272
*Un nuevo tipo de ciencia* (Wolfram) 224, 227
útero 58–59, 203

**V**
valor aleatorio 127
velocidad de la luz 15–16, 19, 269, 305
vía auditiva 92
vía sensitiva 57, 90
vía tacto-sensorial 94
vida humana 265

**W**
Washington University in St. Louis 124
Watson, James D. 8, 14, 273
Watson (ordenador de IBM) xiii, 6, 102, 151–53, 157–64, 169–70, 192, 221, 227, 234, 251, 256–58, 260, 302
Watts, Lloyd xviii, 92, 277

Wedeen, Van J. 77–78, 85, 124, 248, 276
Werblin, Frank S. 90
Whitehead, Alfred North 173
*Whole Brain Emulation: A Roadmap* (Bostrom, Sandberg) 124–26, 279
Wiener, Norbert 110, 138
Wikipedia 6, 150, 159, 163, 168, 220, 256–57, 267, 273
Wittgenstein, Ludwig 209–10
Wolfram Alpha 154, 163, 169

**Y**
Young, Thomas 14

**Z**
zombis 193–94
Zuo, Yi 84, 277
Zuse, Konrad 180